# Model validation, automated variable selection and final model tuning Project

**Purpose:** In this modeling assignment we will finish building linear regression models to predict the home sale price. As such the response variable is: SALEPRICE (Y). We will begin by fitting specific models and looking at diagnostic and model fit information. Models will progressively become more involved and complex over the span of this assignment.

**Data:** The data for this assignment is the Ames, Iowa housing data set. This data is posted in Canvas.

**Explanatory Variables:** All continuous and categorical variables in the AMES Housing data set.

***(1) Preparing the Categorical Variables***

Sample population definition:

```
mydata <-filter(mydata,BldgType == "1Fam")
mydata <-filter(mydata,Zoning %in% c("RH", "RL", "RM", "FV"))
mydata <-filter(mydata,SaleCondition == "Normal")
mydata <- mydata[mydata$GrLivArea<=4000,]
mydata <- mydata[mydata$TotalFloorSF<=4000,]
mydata <- mydata[mydata$SalePrice<=500000,]
```

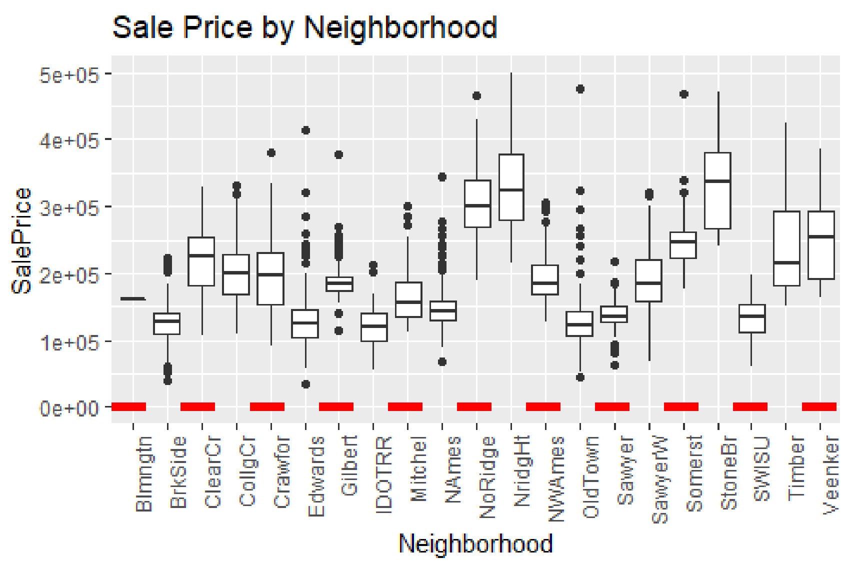
Below are all the categorical variables we will evaluate and selecting the ones to include for the modeling.

```
[1] "Zoning"       "Street"       "LotShape"      "LandContour"   "Utilities"
 [6] "LotConfig"     "LandSlope"     "Neighborhood"  "Condition1"    "Condition2"
[11] "BldgType"      "HouseStyle"    "RoofStyle"     "RoofMat"       "Exterior1"
[16] "Exterior2"     "MasVnrType"    "ExterQual"     "ExterCond"     "Foundation"
[21] "BsmtQual"      "BsmtCond"      "BsmtExposure"  "BsmtFinType1"  "BsmtFinType2"
[26] "Heating"       "HeatingQC"     "CentralAir"    "Electrical"    "KitchenQual"
[31] "Functional"    "GarageType"    "GarageFinish"  "GarageQual"    "GarageCond"
[36] "PavedDrive"    "SaleType"      "SaleCondition"
```

Using some high-level intuition, I sampled majority of the categorical variables and reporting the R-squared value to understand the final ones that I would consider retaining. Now let's look closer at the 5 highlighted in the below table and finalize our decision.

| lm(SalePrice~) | R-Squared Value |
|---|---|
| Neighborhood | 60.22% |
| RoofStyle | 3.18% |
| RoofMat | 1.55% |
| MasVnrType | 19.52% |
| Foundation | 27.33% |
| Heating | 0.78% |
| Zoning | 10.62% |
| ExterQual | 47.16% |
| HouseStyle | 12.87% |
| Exterior1 | 21.64% |
| Exterior2 | 21.95% |
| KitchenQual | 43.88% |
| GarageType | 18.91% |
| GarageQual | 4.37% |
| GarageFinish | 27.60% |
| BsmtQual | 47.93% |
| BsmtFinType1 | 24.62% |
| BsmtFinType2 | 0.90% |
| Bsmt Cond | 2.50% |
| BsmtExposure | 17.18% |
| PavedDrive | 8.40% |
| LandContour | 2.89% |
| Street | 0.03% |
| LotShape | 8.38% |
| LotConfig | 1.15% |
| SaleType | 1.49% |
| Condition1 | 5.18% |
| Condition2 | 1.56% |

1. Neighborhood
   Lm() function displayed the statistical significance by individual level. It seems only 3 out of all neighborhoods shows the significance. They are North Ridge, N. Ridge Height, and Stone Bridge. Boxplots further confirmed there are significant outliers within majority of the neighborhoods with their relationship to the SalePrice. This may not be a good categorical variable to include for our modeling.



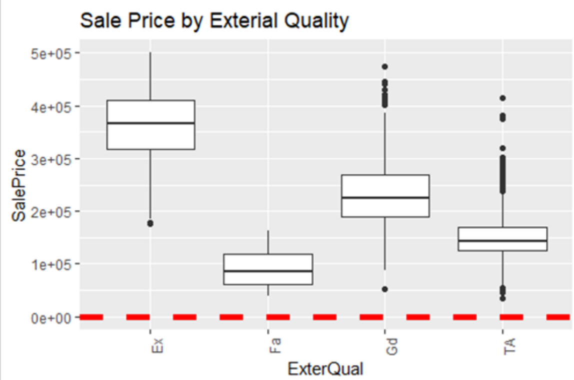

```
Call:
lm(formula = SalePrice ~ Neighborhood, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-123008  -24004   -4654   19057  346448

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            159895      43388   3.685 0.000235 ***
NeighborhoodBrkSide    -33155      43614  -0.760 0.447231
NeighborhoodClearCr     58506      43971   1.331 0.183485
NeighborhoodCollgCr     39884      43490   0.917 0.359207
NeighborhoodCrawfor     39126      43665   0.896 0.370335
NeighborhoodEdwards    -26939      43556  -0.618 0.536326
NeighborhoodGilbert     29315      43557   0.673 0.501018
NeighborhoodIDOTRR     -38787      43820  -0.885 0.376187
NeighborhoodMitchel      6632      43649   0.152 0.879248
NeighborhoodNAmes      -12991      43448  -0.299 0.764966
NeighborhoodNoRidge    150811      43726   3.449 0.000575 ***
NeighborhoodNridgHt    173745      43726   3.974 7.34e-05 ***
NeighborhoodNWAmes      34489      43580   0.791 0.428804
NeighborhoodOldTown    -31343      43511  -0.720 0.471397
NeighborhoodSawyer     -22569      43567  -0.518 0.604498
NeighborhoodSawyerW     30613      43631   0.702 0.482989
NeighborhoodSomerst     88623      43711   2.027 0.042748 *
NeighborhoodStoneBr    173315      45160   3.838 0.000128 ***
NeighborhoodSWISU      -26911      44022  -0.611 0.541060
NeighborhoodTimber      82100      43829   1.873 0.061188 .
NeighborhoodVeenker     92596      44646   2.074 0.038209 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43390 on 1960 degrees of freedom
Multiple R-squared:  0.6022,	Adjusted R-squared:  0.5981
F-statistic: 148.3 on 20 and 1960 DF,  p-value: < 2.2e-16
```

| | Neighborhood | SalePrice |
|---|---|---|
| 1 | Blmngtn | 159895.0 |
| 2 | BrkSide | 126740.4 |
| 3 | ClearCr | 218400.9 |
| 4 | CollgCr | 199779.2 |
| 5 | Crawfor | 199021.4 |
| 6 | Edwards | 132956.2 |
| 7 | Gilbert | 189209.6 |
| 8 | IDOTRR | 121108.1 |
| 9 | Mitchel | 166527.1 |
| 10 | NAmes | 146903.7 |
| 11 | NoRidge | 310705.6 |
| 12 | NridgHt | 333639.8 |
| 13 | NWAmes | 194384.1 |
| 14 | OldTown | 128551.8 |
| 15 | Sawyer | 137326.1 |
| 16 | SawyerW | 190508.2 |
| 17 | Somerst | 248517.9 |
| 18 | StoneBr | 333210.0 |
| 19 | SWISU | 132983.8 |
| 20 | Timber | 241995.2 |
| 21 | Veenker | 252491.2 |

2. Exterior Quality

Mean deviation and boxplot showing we should consider this categorical variable in explaining SalePrice. This variable alone can explain 47.16% of the SalePrice variance. All levels within the categorical variable shows statistical significance.

```
Call:
lm(formula = SalePrice ~ ExterQual, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-181491  -29343   -6593   25157  266157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   357991       7873   45.47   <2e-16 ***
ExterQualFa  -266551      14728  -18.10   <2e-16 ***
ExterQualGd  -124935       8128  -15.37   <2e-16 ***
ExterQualTA  -209149       7991  -26.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49790 on 1977 degrees of freedom
Multiple R-squared:  0.4716,    Adjusted R-squared:  0.4708
F-statistic: 588.1 on 3 and 1977 DF,  p-value: < 2.2e-16
```

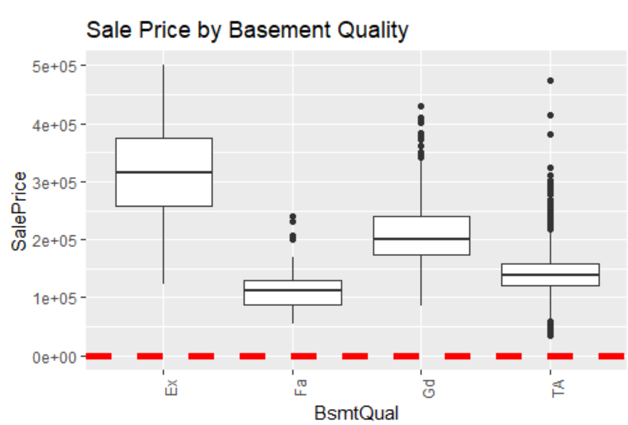



| | ExterQual | SalePrice |
|---|---|---|
| 1 | Ex | 357991.25 |
| 2 | Fa | 91440.62 |
| 3 | Gd | 233055.86 |
| 4 | TA | 148842.80 |

3. Basement Quality

Mean deviation and boxplot showing we should consider this categorical variable in explaining SalePrice. This variable alone can explain 47.93% of the SalePrice variance. All levels within the categorical variable shows statistical significance.

```
Call:
lm(formula = SalePrice ~ BsmtQual, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-191441  -29044   -6566   21556  331434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   314941       4531   69.51   <2e-16 ***
BsmtQualFa   -200077       7623  -26.25   <2e-16 ***
BsmtQualGd   -105998       4868  -21.77   <2e-16 ***
BsmtQualTA   -171376       4786  -35.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 49430 on 1977 degrees of freedom
Multiple R-squared:  0.4793,    Adjusted R-squared:  0.4785
F-statistic: 606.5 on 3 and 1977 DF,  p-value: < 2.2e-16
```

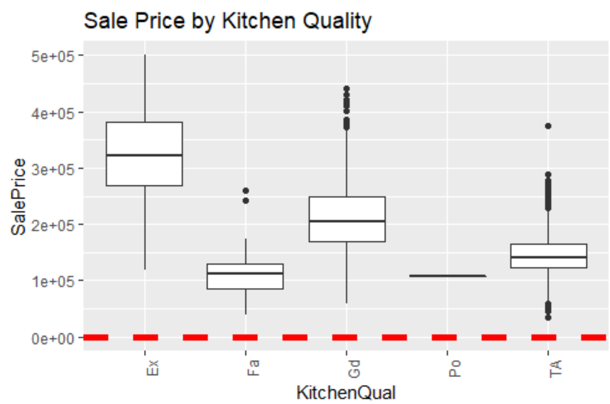


BsmtQual SalePrice
1 Ex 314941.3
2 Fa 114864.6
3 Gd 208943.7
4 TA 143565.7

4. Kitchen Quality

Mean deviation and boxplot showing we should consider this categorical variable in explaining SalePrice. This variable alone can explain 43.88% of the SalePrice variance.

```
Call:
lm(formula = SalePrice ~ KitchenQual, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-200320  -29856   -5027   25973  230173

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     319320       5410  59.023  < 2e-16 ***
KitchenQualFa  -208105       8996 -23.134  < 2e-16 ***
KitchenQualGd  -107563       5713 -18.827  < 2e-16 ***
KitchenQualPo  -211820      51609  -4.104 4.22e-05 ***
KitchenQualTA  -174293       5636 -30.928  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51320 on 1976 degrees of freedom
Multiple R-squared:  0.4388,    Adjusted R-squared:  0.4377
F-statistic: 386.2 on 4 and 1976 DF,  p-value: < 2.2e-16
```




KitchenQual SalePrice
1 Ex 319319.5
2 Fa 111214.7
3 Gd 211756.1
4 Po 107500.0
5 TA 145026.9

```
lm(formula = SalePrice ~ KitchenQual + BsmtQual + ExterQual,
    data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-117987  -26711   -3711   20739  253324

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     383458       6733  56.953  < 2e-16 ***
KitchenQualFa   -97482       8443 -11.546  < 2e-16 ***
KitchenQualGd   -51053       5854  -8.721  < 2e-16 ***
KitchenQualPo  -140277      42317  -3.315 0.000933 ***
KitchenQualTA   -77018       6106 -12.613  < 2e-16 ***
BsmtQualFa     -111721       7236 -15.439  < 2e-16 ***
BsmtQualGd      -52023       4824 -10.784  < 2e-16 ***
BsmtQualTA      -87071       5164 -16.859  < 2e-16 ***
ExterQualFa    -126982      13546  -9.374  < 2e-16 ***
ExterQualGd     -50726       8209  -6.179 7.82e-10 ***
ExterQualTA     -83658       8565  -9.768  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41850 on 1970 degrees of freedom
Multiple R-squared:  0.628,	Adjusted R-squared:  0.6261
F-statistic: 332.6 on 10 and 1970 DF,  p-value: < 2.2e-16
```

All individual levels within the categorical variables selected here show statistical significance individually in contributing to the SalePrice.

Creating the dummy variables for the 3 prior to our modeling framework.

new<- dummy.data.frame(cleandata, names = c("KitchenQual","BsmtQual","ExterQual") , sep = ".")

Combined model for running LM formula, it produced R-Squared value at 62.8% in explaining the variance of the SalePrice. More deviations among the mean among levels within the categorical variable, then it shows the ability in explaining the response variable, in this case, SalePrice.

For our predictive modeling framework, I will be focusing on 3 categorical variables: ExterQual, KitchenQual, and BsmtQual.

### *(2) The Predictive Modeling Framework*

Below is a table of observation counts for the train/test data partition.

|DataFrame | ObsCounts| PercentOfObs|
|:---------|---------:|------------:|
|train.df  |      1392|    0.7026754|
|test.df   |       589|    0.2973246|

### *(3) Model Identification by Automated Variable Selection*

Create a pool of candidates. Using correlation table and only reviewing top 10 and combining the 3 categorical variables. Below are our total 24 variables, which we will use typical Kitchen Quality, Basement Quality and Exterior Quality as the control dummy variables.

| rowname | SalePrice |
|---|---|
| OverallQual | 0.801761801 |
| TotalFloorSF | 0.782481418 |
| GrLivArea | 0.77535809 |
| GarageCars | 0.660977076 |
| TotalBsmtSF | 0.650020052 |
| GarageArea | 0.640403126 |
| FullBath | 0.609247762 |
| TotRmsAbvGrd | 0.599288489 |
| SF | 0.558873904 |
| MasVnrArea | 0.554924441 |

| | Variable Original | Transformed |
|---|---|---|
| 1 | 'SalePrice' | |
| 2 | 'QualityIndex' | |
| 3 | 'TotalSqftCalc' | |
| 4 | 'YrSold' | |
| 5 | 'FullBath' | |
| 6 | 'GarageArea' | |
| 7 | 'GarageCars' | |
| 8 | TotRmsAbvGrd' | |
| 9 | 'LotArea' | |
| 10 | MasVnrArea' | |
| 11 | 'WoodDeckSF' | |
| 12 | BsmtQual | BsmtQual.Ex' |
| 13 | | 'BsmtQual.Fa' |
| 14 | | 'BsmtQual.Gd' |
| 15 | | BsmtQual.TA' |
| 16 | KitchenQual | 'KitchenQual.Fa' |
| 17 | | 'KitchenQual.Gd' |
| 18 | | KitchenQual.Ex' |
| 19 | | 'KitchenQual.Po' |
| 20 | | 'KitchenQual.TA' |
| 21 | ExterQual | 'ExterQual.Ex' |
| 22 | | 'ExterQual.Fa' |
| 23 | | 'ExterQual.Gd' |
| 24 | | ExterQual.TA' |

***Model Identification:***

Variables selected in my model do not present collinear relationship. But in the junk model we do. Since Quality index was calculated based on:

mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond

Running initial model comparison for forward backward and stepwise.

| Model Comparison | R-Sqaured | Adjusted R-Sqaured | RSE | F-Statistic |
|---|---|---|---|---|
| forward.lm | 0.896 | 0.8947 | 22440 | 696.2 |
| backward.lm | 0.8959 | 0.8947 | 22440 | 739.9 |
| stepwise.lm | 0.896 | 0.8947 | 22440 | 696.2 |
| jumk.lm | 0.8352 | 0.8346 | 28130 | 1405 |

Compute the VIF values for the variable selection models. Using 10 as the VIF threshold for the validation, I do not see any variables in three models exceed 6. However, based on the issue of collinearity in the junk model, Quality Index, Overall Quality, and Overall Condition violated that and indicating when including variables bares this relationship, they should be dropped from initial model and a re-evaluation is needed.

| | Variables | fwd VIF | bkw VIF | step VIF | jnk VIP |
|---|---|---|---|---|---|
| 1 | GarageCars | 5.076035 | 5.075905 | 5.076035 | NA |
| 2 | GarageArea | 4.584385 | 4.578257 | 4.584385 | NA |
| 3 | ExterQual.TA | 2.348591 | 2.308739 | 2.348591 | NA |
| 4 | Total SqftCalc | 2.309327 | 2.220551 | 2.309327 | 2.581736 |
| 5 | FullBath | 2.143889 | 2.141817 | 2.143889 | NA |
| 6 | KitchenQual.Ex | 2.07072 | 2.057533 | 2.07072 | NA |
| 7 | BsmtQual.Gd | 1.955835 | 1.955823 | 1.955835 | NA |
| 8 | BsmtQual.Ex | 1.93763 | 1.937449 | 1.93763 | NA |
| 9 | TotRmsAbvGrd | 1.931951 | 1.931929 | 1.931951 | NA |
| 10 | KitchenQual.Gd | 1.887583 | 1.87688 | 1.887583 | NA |
| 11 | ExterQual.Ex | 1.676307 | 1.887573 | 1.676307 | NA |
| 12 | QualityIndex | 1.382491 | 1.356689 | 1.382491 | 30.83422 |
| 13 | MasVnrArea | 1.369209 | 1.368952 | 1.369209 | NA |
| 14 | WoodDeckSF | 1.166341 | 1.165658 | 1.166341 | NA |
| 15 | ExterQual.Fa | 1.160879 | NA | 1.160879 | NA |
| 16 | LotArea | 1.109707 | 1.109566 | 1.109707 | NA |
| 17 | BsmtQual.Fa | 1.080743 | 1.052942 | 1.080743 | NA |
| 18 | OverallQual | NA | NA | NA | 18.1801 |
| 19 | OverallCond | NA | NA | NA | 16.47569 |
| 20 | GrLivArea | NA | NA | NA | 2.924638 |

Another lesson here worth noting, we can't simply rely on the statistical significance when evaluating predictors, since junk model's coefficients at individual level all has a high t-value. This can be misleading if the modeler is relying on t-value and p value alone.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.412e+05  1.453e+04  -9.719  < 2e-16 ***
OverallQual    3.575e+04  2.443e+03  14.633  < 2e-16 ***
OverallCond    9.923e+03  2.608e+03   3.805 0.000148 ***
QualityIndex  -1.827e+03  4.467e+02  -4.089 4.58e-05 ***
GrLivArea      2.261e+01  2.681e+00   8.432  < 2e-16 ***
TotalSqftCalc  3.926e+01  1.754e+00  22.379  < 2e-16 ***
```

***Model Comparison:***

Based on comparison by adding AIC, BIC, MAE, and MSE. Below is the summary result. If looking closer, ranking is not consistent.

First three models all have the same adjusted R-squared value, so they all can explain the model variance with 89.6%.

Looking at AIC and BIC, backward model stood up, and that's important since both metrics adds penalty if additional variables are included in the model.

We certainly want lowest MAE and MSE, but backward model which has the lowest AIC and BIC has slightly larger MAE and MSE value than forward and stepwise models.

Noting forward and stepwise models have the identical results. In this case, I do not have ranking consistently the same, and it truly depends on which one I want to prioritize in selecting the final model.

| Model Comparison | Adj. R-Sqaured | Rank | AIC | Rank | BIC | Rank | MAE | Rank | MSE | Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| forward.lm | 0.8947 | 1 | 31862.25 | 2 | 31961.78 | 2 | 16742.33 | 1 | 497133799 | 1 |
| backward.lm | 0.8947 | 1 | 31860.84 | 1 | 31955.13 | 1 | 16748.89 | 2 | 497343870 | 2 |
| stepwise.lm | 0.8947 | 1 | 31862.25 | 2 | 31961.78 | 2 | 16742.33 | 1 | 497133799 | 1 |
| jumk.lm | 0.8346 | 2 | 32479.06 | 3 | 32515.73 | 3 | 20714.09 | 3 | 787773706 | 3 |

**(4) *Predictive Accuracy***

| Model Comparison | MAE | Rank | MSE | Rank |
|---|---|---|---|---|
| forward.lm | 17554.14 | 2 | 594463073 | 2 |
| backward.lm | 17539.27 | 1 | 593756604 | 1 |
| stepwise.lm | 17554.14 | 2 | 594463073 | 2 |
| jumk.lm | 20552.3 | 3 | 751329963 | 3 |

Prediction accuracy against test data shows backward model, or #2 is slightly better even though the MAE/MSE previously was slightly less desirable than forward and stepwise model. I think this indicates AIC and BIC are more important fitness to consider.

**(5) *Operational Validation***

Define a variable called PredictionGrade, and consider the predicted value to be 'Grade 1' if it is within ten percent of the actual value, 'Grade 2' if it is not Grade 1 but within fifteen percent of the actual value, Grade 3 if it is not Grade 2 but within twenty-five percent of the actual value, and 'Grade 4' otherwise.

Side by side comparison of in-sampling training data and out-of-sample.

| Training | | | | Testing | | | |
|---|---|---|---|---|---|---|---|
| forward.PredictionGrade | | | | forward.testPredictionGrade | | | |
| Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] | Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] |
| 0.61566092 | 0.18606322 | 0.14008621 | 0.05818966 | 0.59762309 | 0.16129032 | 0.17657046 | 0.06451613 |
| | | | | | | | |
| backward.PredictionGrade | | | | backward.testPredictionGrade | | | |
| Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] | Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] |
| 0.61494253 | 0.18606322 | 0.14080460 | 0.05818966 | 0.59762309 | 0.16129032 | 0.17826825 | 0.06281834 |
| | | | | | | | |
| stepwise.PredictionGrade | | | | stepwise.testPredictionGrade | | | |
| Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] | Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] |
| 0.61566092 | 0.18606322 | 0.14008621 | 0.05818966 | 0.59762309 | 0.16129032 | 0.17657046 | 0.06451613 |
| | | | | | | | |
| junk.PredictionGrade | | | | junk.testPredictionGrade | | | |
| Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] | Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] |
| 0.5316092 | 0.1752874 | 0.1716954 | 0.1214080 | 0.5314092 | 0.1680815 | 0.1765705 | 0.1239389 |

Model accuracy is decreasing between training and testing for forward, backward and stepwise across Grade 1 through Grade 4. Junk model however shows a very small degradation comparing to the top 3 models. General statement is the predicted value is higher when the data is closer to the actual value within 10%.

| | Training | | | | Testing | | | |
|---|---|---|---|---|---|---|---|---|
| | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Grade 1 | Grade 2 | Grade 3 | Grade 4 |
| forward.PredictionGrade | 1 | 1 | 3 | 2 | 1 | 2 | 3 | 2 |
| backward.PredictionGrade | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 3 |
| stepwsie.PredictionGrade | 1 | 1 | 3 | 2 | 1 | 2 | 3 | 2 |
| junk.PredictionGrade | 3 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |

Ranking here is more thorough than the way of evaluating the model above just using just MSE and MAE. This would make more sense for business interpretation. If specifically looking at three models excluding junk. Backward model showing best predicting in test in Grade 3. Otherwise, it's the same comparing to forward and stepwise models in Grade 1 and 2. But least predicting when looking at Grade 4.

Using the prediction grade method, I noticed junk model is predicting much better in test and ranked 1 in Grade 4.

The GSEs (Fannie Mae and Freddie Mac) rate an AVM model as 'underwriting quality', if the model is accurate to within ten percent more than fifty percent of the time. Then all 4 models can be considered as 'underwriting quality'.

6) ***Final model selection***

Based on the prediction grade evaluation, I am picking forward.lm/stepwise.lm model(pretty consistent in terms of performance). We will use this one as the starting point to review and tune to get to our final model.

Examine the impact of removing each variable from the final model, one at a time. Comparing to the total variables selected originally, "YrSold" was eliminated in the variable selection. I will use the R-Squared value as the benchmark before starting to remove one variable at the time and compare the difference based on R-Squared value.

First let's look at the benchmark model and develop a plan in terms of the sequence in reducing the variables.

```
Call:
lm(formula = SalePrice ~ QualityIndex + TotalSqftCalc + FullBath +
    GarageArea + GarageCars + TotRmsAbvGrd + LotArea + MasVnrArea +
    WoodDeckSF + BsmtQual.Ex + BsmtQual.Fa + BsmtQual.TA + BsmtQual.Gd +
    KitchenQual.Fa + KitchenQual.Gd + KitchenQual.Po + KitchenQual.TA +
    KitchenQual.Ex + ExterQual.Ex + ExterQual.Fa + ExterQual.Gd +
    ExterQual.TA, data = train.clean1)

Residuals:
   Min     1Q Median     3Q    Max
-90210 -13180    589  12633 102371

Coefficients: (4 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.555e+04  6.062e+03   5.864 5.66e-09 ***
QualityIndex    1.120e+03  7.554e+01  14.832  < 2e-16 ***
TotalSqftCalc   3.216e+01  1.324e+00  24.289  < 2e-16 ***
FullBath        5.532e+03  1.635e+03   3.383 0.000736 ***
GarageArea      1.675e+01  6.452e+00   2.596 0.009519 **
GarageCars      9.230e+03  1.905e+03   4.844 1.42e-06 ***
TotRmsAbvGrd    3.835e+03  5.929e+02   6.469 1.37e-10 ***
LotArea         9.252e-01  8.640e-02  10.708  < 2e-16 ***
MasVnrArea      4.357e+01  4.364e+00   9.983  < 2e-16 ***
WoodDeckSF      1.301e+01  4.869e+00   2.671 0.007643 **
BsmtQual.Ex     3.523e+04  3.222e+03  10.935  < 2e-16 ***
BsmtQual.Fa    -2.109e+04  3.587e+03  -5.878 5.20e-09 ***
BsmtQual.TA    -1.536e+04  1.728e+03  -8.890  < 2e-16 ***
BsmtQual.Gd            NA         NA      NA       NA
KitchenQual.Fa -3.275e+04  5.392e+03  -6.074 1.61e-09 ***
KitchenQual.Gd -2.651e+04  3.864e+03  -6.862 1.03e-11 ***
KitchenQual.Po         NA         NA      NA       NA
KitchenQual.TA -3.605e+04  4.074e+03  -8.849  < 2e-16 ***
KitchenQual.Ex         NA         NA      NA       NA
ExterQual.Ex    3.873e+04  5.355e+03   7.233 7.84e-13 ***
ExterQual.Fa   -5.397e+03  6.492e+03  -0.831 0.405935
ExterQual.Gd    1.614e+04  1.951e+03   8.271 3.10e-16 ***
ExterQual.TA           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22440 on 1373 degrees of freedom
Multiple R-squared:  0.896,	Adjusted R-squared:  0.8947
F-statistic: 657.5 on 18 and 1373 DF,  p-value: < 2.2e-16
```

We will first reduce Garage Area since it has the lowest t-value and closest to 0. Using this logic, we will go through one by one and monitoring the R-Squared changes.

We were going through the evaluation and until we reached 5.28% changes when removing QualityIndex from the model. I decided to keep this variable back to the model. Then removing BsmtQual dummy variables, and it reduced the R-Squared value by 12.26%. That’s a significant change, as a result, I am keeping BsmtQual dummy variables back to the model.

| | Model | R-Sqaured | Delta | |
|---|---|---|---|---|
| Baseline | forward.lm | 89.60% | | |
| Reduce | GarageArea | 89.55% | -0.05% | |
| Reduce | WoodDeckSF | 89.50% | -0.05% | |
| Reduce | FullBath | 89.41% | -0.09% | |
| Reduce | GarageCars | 88.30% | -1.11% | |
| Reduce | TotRmsAbvGrd | 87.68% | -0.62% | |
| Reduce | LotArea | 86.64% | -1.04% | |
| Reduce | MasVnrArea | 85.64% | -1.00% | |
| Reduce | ExterQual All | 84.36% | -1.28% | |
| Reduce | KitchenQual All | 82.15% | -2.21% | |
| Reduce | QualityIndex | 76.87% | -5.28% | Add it back |
| Reduce | BasementQual All | 69.89% | -12.26% | Add it back |

Based on the iteration above, we landed at our final model. Which is QualityIndex, TotalSqftCalc and BasmentQaul related categorical dummy variables and rerun the model to see the coefficients.

```
Call:
lm(formula = SalePrice ~ QualityIndex + TotalSqftCalc + BsmtQual.Ex +
    BsmtQual.Fa + BsmtQual.TA + BsmtQual.Gd, data = train.clean1)

Residuals:
    Min      1Q  Median      3Q     Max
-136507  -17806      -7   15822  130975

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    29350.21    3937.36   7.454 1.58e-13 ***
QualityIndex    1819.30      89.89  20.240  < 2e-16 ***
TotalSqftCalc     52.14       1.34  38.920  < 2e-16 ***
BsmtQual.Ex    67445.96    3566.77  18.910  < 2e-16 ***
BsmtQual.Fa   -42934.13    4422.33  -9.708  < 2e-16 ***
BsmtQual.TA   -36437.89    1773.54 -20.545  < 2e-16 ***
BsmtQual.Gd          NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29270 on 1386 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8208
F-statistic:  1275 on 5 and 1386 DF,  p-value: < 2.2e-16
```

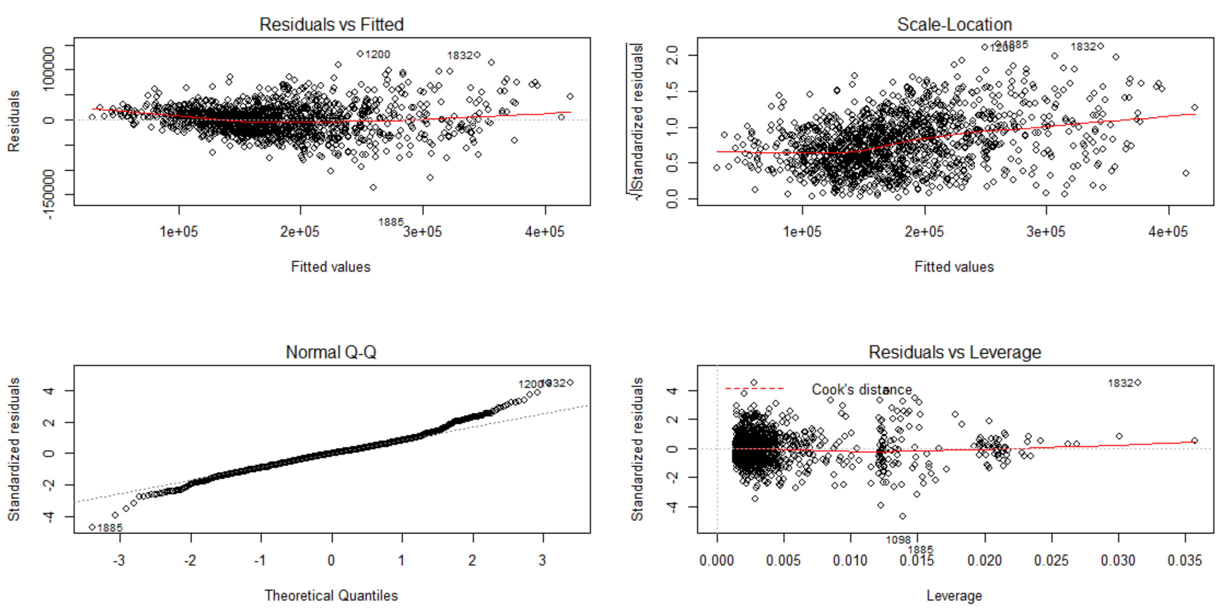
$y=29350.21+1819.30B1+52.14B2+67445.96B3-42934.13B4-36437.89B5$

- SalePrice will be $29,350.21 if everything is 0, and basement quality is great as the basis of interpretation. Which is low since there would not be a house with 0 total square footage for sale.
- Per quality index change, it will add $1819.30 to the SalePirce
- Per total square footage change will add $52.14
- If the basement quality is excellent, it will add $67445.96

- If the basement quality is fair, it will reduce $42934.13
- If the basement quality is typical, it will reduce $36437.89
- We reject null hypothesis for B1…B5. T- value and p-value are indicator they are significant to the model
- And the overall model at F (3, 1386)=1275, which we reject the null hypothesis under the Omibus, and it is statistically significant in explaining the variance, which it explains the model at 82.15%.
- QQ plot shows normality except around the tails due to extreme outliers. (1885, 1832, 1200)
- Residual and Fitted graph show a slight parabola shape, so we can say the model does not meet the homoscedasticity assumption since the residuals are not equally spread around the y = 0 line. linear model assumption is there since the redline through our scatterplot is fairly straight.
- Scale location graph is supporting the evaluation of homoscedasticity. We see the red line is sloping slightly up and the data points are randomly spread out. So, this model violated the assumption.
- Cook's distance shows other than few extreme values, we see most within the range.

Running prediction grade against the final model by validating against out-of-sample data. We can say the model is accurate within 10% more than 50% of the time.

```
final.testPredictionGrade
  Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
       0.5110357          0.1731749          0.2020374          0.1137521
```




```
      rstudent unadjusted p-value Bonferroni p
1885 -4.731801          2.4530e-06    0.0034146
1832  4.552317          5.7711e-06    0.0080333
1200  4.511422          6.9843e-06    0.0097221
```

**ANOVA TESTING**

However, we need to continue with modeling since we are using a categorical value in addition to numerical, therefore we need to test the unequal slope.

Since TotalSqftCalc is using basement finished sf a part of the calculation, let's add the effect variable and see if we notice any improvement to our R-Squared value and evaluate the interaction between TotalSqftCalc and BasmentQuality.

cleandata$TotalSqftCalc <- cleandata$BsmtFinSF1+cleandata$BsmtFinSF2+cleandata$GrLivArea

```
Call:
lm(formula = SalePrice ~ QualityIndex + TotalSqftCalc + TotalSqftCalc *
    BsmtQual.Ex + TotalSqftCalc * BsmtQual.Fa + TotalSqftCalc *
    BsmtQual.TA, data = train.clean1)

Residuals:
    Min      1Q  Median      3Q     Max
-112924  -17599     -63   15649  137450

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              28256.089   4975.355   5.679 1.65e-08 ***
QualityIndex              1847.910     88.850  20.798  < 2e-16 ***
TotalSqftCalc               52.175      1.890  27.612  < 2e-16 ***
BsmtQual.Ex              -2709.778  13301.024  -0.204   0.8386
BsmtQual.Fa             -24640.486  12511.115  -1.969   0.0491 *
BsmtQual.TA             -27941.950   5630.715  -4.962 7.83e-07 ***
TotalSqftCalc:BsmtQual.Ex   24.410      4.637   5.264 1.63e-07 ***
TotalSqftCalc:BsmtQual.Fa  -12.890      8.065  -1.598   0.1102
TotalSqftCalc:BsmtQual.TA   -4.802      2.716  -1.768   0.0773 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28880 on 1383 degrees of freedom
Multiple R-squared:  0.8267,    Adjusted R-squared:  0.8257
F-statistic: 824.5 on 8 and 1383 DF,  p-value: < 2.2e-16
```

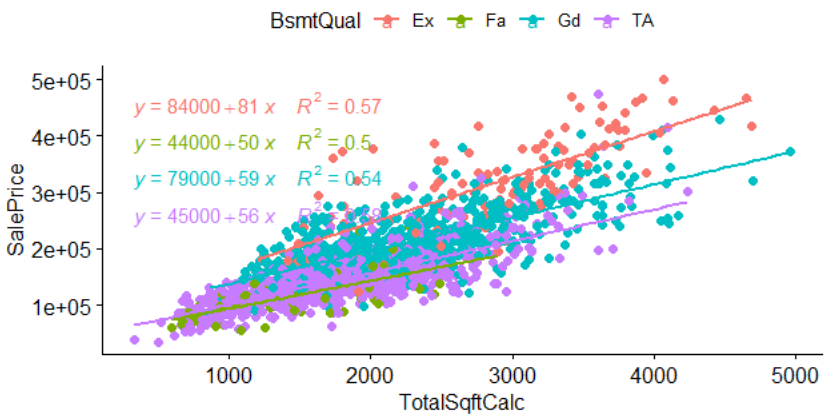
```
Analysis of Variance Table

Model 1: SalePrice ~ QualityIndex + TotalSqftCalc + BsmtQual.Ex + BsmtQual.Fa +
    BsmtQual.TA + TotalSqftCalc * BsmtQual.Ex + TotalSqftCalc *
    BsmtQual.Fa + TotalSqftCalc * BsmtQual.TA
Model 2: SalePrice ~ QualityIndex + TotalSqftCalc + BsmtQual.Ex + BsmtQual.Fa +
    BsmtQual.TA
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1   1383 1.1531e+12
2   1386 1.1878e+12 -3 -3.4687e+10 13.867 6.528e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model testing effect of TotalSqftCalc is the full model (unequal slope) and nesting our final model as the reduced model. Using anova () we can validate the F-Statistics between the full and the nested model.

y=29256.089+1847.91B1+52.175B2-2709.778B3-24640.486B4-27941.95B5+24.41B6-12.89B7-4.802B8

- SalePrice will be $29,256.089 if everything is 0, and basement quality is great as the basis of interpretation. Which is low since there would not be a house with 0 total square footage for sale.
- Per quality index change, it will add $1847.91 to the SalePrice; per total square footage change will add $52.175. Change to quality of basement from great to excellent will decrease 2709.778, and change to fair will decrease 24640.486, and change to typical will decrease 27941.95. interaction of totalsqftcalc and quality of basement in excellent can add 24.10, interaction of totalsqftcalc and quality of basement in fair can decrease 12.89, and interaction of totalsqftcalc and quality of basement in typical can decrease 4.802 to the sale price.
- We reject null hypothesis for B1,B2, B5, and B6. T- value and p-value are indicator they are significant to the model
- We fail to reject null hypothesis for B3, B4, B7, and B8. Their t-value and p-value indicate they are not significant to the model.

- F-statistics (8,1383) =824.5 with p value <2.2e-16. This model is significant and ale to explain 82.67% to the variance of SalePrice.
- Nesting model evaluation using F-statistics and anova() function. F value is 13.867 with p-value at 6.528e-09. We fail to reject null hypothesis that unequal slope parameters being zeros, there is interaction between TotalSqftCalc and BsmtQual.




Next looking at the interaction between BsmtQual with QualityIndex.

```
Call:
lm(formula = SalePrice ~ QualityIndex + TotalSqftCalc + QualityIndex *
    BsmtQual.Ex + QualityIndex * BsmtQual.Fa + QualityIndex *
    BsmtQual.TA, data = train.clean1)

Residuals:
    Min      1Q  Median      3Q     Max
-119125  -17391    -374   15078  148738

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             14308.68    6413.52   2.231   0.0258 *
QualityIndex             2341.41     176.98  13.230  < 2e-16 ***
TotalSqftCalc              50.55       1.30  38.878  < 2e-16 ***
BsmtQual.Ex           -163839.18   25987.99  -6.304 3.88e-10 ***
BsmtQual.Fa            -15667.97   11688.88  -1.340   0.1803
BsmtQual.TA             -9707.47    7155.09  -1.357   0.1751
QualityIndex:BsmtQual.Ex  5527.34    628.39   8.796  < 2e-16 ***
QualityIndex:BsmtQual.Fa  -863.02    354.61  -2.434   0.0151 *
QualityIndex:BsmtQual.TA  -794.09    201.29  -3.945 8.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28150 on 1383 degrees of freedom
Multiple R-squared:  0.8353,	Adjusted R-squared:  0.8344
F-statistic: 876.9 on 8 and 1383 DF,  p-value: < 2.2e-16
```

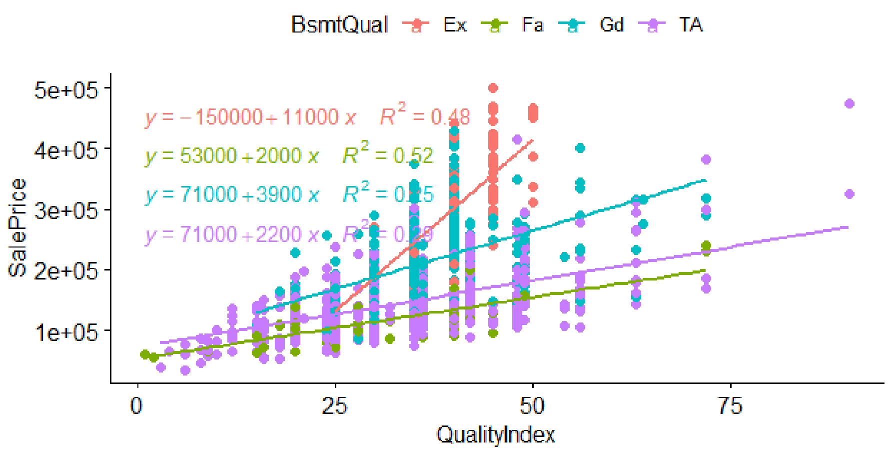
```
Model 1: SalePrice ~ QualityIndex + TotalSqftCalc + QualityIndex * BsmtQual.Ex +
    QualityIndex * BsmtQual.Fa + QualityIndex * BsmtQual.TA
Model 2: SalePrice ~ QualityIndex + TotalSqftCalc + BsmtQual.Ex + BsmtQual.Fa +
    BsmtQual.TA
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   1383 1.0956e+12
2   1386 1.1878e+12 -3 -9.2225e+10 38.806 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model testing effect of QualityIndex is the full model (unequal slope) and nesting our final model as the reduced model. Using anova () we can validate the F-Statistics between the full and the nested model.

y=14308.68+2341.41B1+50.55B2-16839.18B3-15667.97B4-9707.47B5+5527.34B6-863.02B7-794.09B8

- SalePrice will be $14308.68 if everything is 0, and basement quality is great as the basis of interpretation.

- Per quality index change, it will add $2341.41 to the SalePrice; per total square footage change will add $50.55. Change to quality of basement from great to excellent will decrease 163839.18, and change to fair will decrease 15667.97 and change to typical will decrease 9707.47. interaction of QualityIndex and quality of basement in excellent condition can add 5527.34, interaction of QualityIndex and quality of basement in fair can decrease 863.02, and interaction of QualityIndex and quality of basement in typical can decrease 794.09 to the sale price.
- We reject null hypothesis for B1,B2, B3, B6 and B8. T- value and p-value are indicator they are significant to the model
- We fail to reject null hypothesis for B4, B5, B7. Their t-value and p-value indicate they are not significant to the model.
- F-statistics (8,1383) =876.9 with p value <2.2e-16. This model is significant and ale to explain 83.53% to the variance of SalePrice.
- Nesting model evaluation using F-statistics and anova() function. F value is 38.806 with p-value at 2.2e-16. We fail to reject null hypothesis that unequal slope parameters being zeros, there is interaction between QualityIndex and BsmtQual.

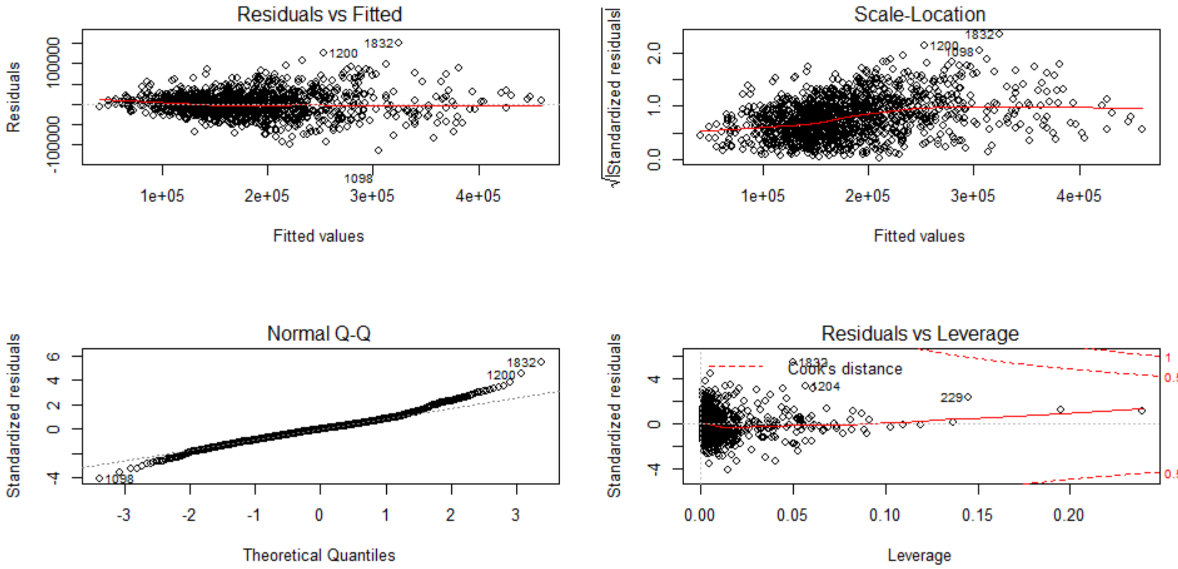


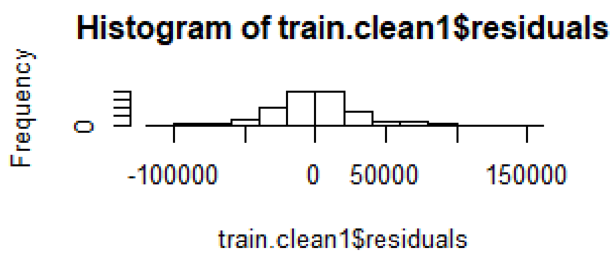

Since both numerical variables interact with basement quality. I am rerunning the final model. BsmtQual.Gr is the basis for interpretation.

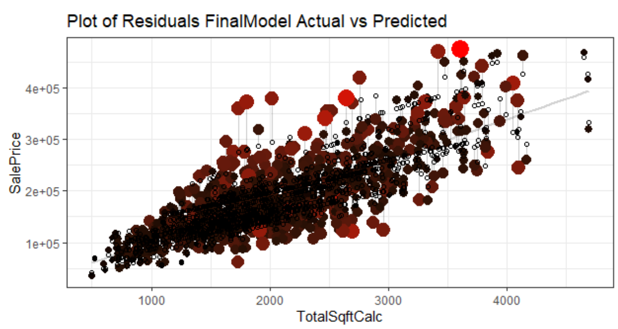
```
Call:
lm(formula = SalePrice ~ QualityIndex + TotalSqftCalc + TotalSqftCalc *
    BsmtQual.Ex + TotalSqftCalc * BsmtQual.Fa + TotalSqftCalc *
    BsmtQual.TA + QualityIndex * BsmtQual.Ex + QualityIndex *
    BsmtQual.Fa + QualityIndex * BsmtQual.TA, data = train.clean1)

Residuals:
    Min      1Q  Median      3Q     Max
-114832  -17233    -517   14952  150792

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.438e+04  6.590e+03   2.183 0.029218 *
QualityIndex               2.343e+03  1.817e+02  12.896  < 2e-16 ***
TotalSqftCalc              5.049e+01  1.916e+00  26.352  < 2e-16 ***
BsmtQual.Ex               -1.722e+05  2.616e+04  -6.583 6.54e-11 ***
BsmtQual.Fa               -5.784e+03  1.455e+04  -0.398 0.690953
BsmtQual.TA               -7.326e+03  7.891e+03  -0.928 0.353345
TotalSqftCalc:BsmtQual.Ex  1.271e+01  4.828e+00   2.633 0.008552 **
TotalSqftCalc:BsmtQual.Fa -9.340e+00  8.192e+00  -1.140 0.254386
TotalSqftCalc:BsmtQual.TA -1.801e+00  2.754e+00  -0.654 0.513218
QualityIndex:BsmtQual.Ex   4.852e+03  6.716e+02   7.224 8.34e-13 ***
QualityIndex:BsmtQual.Fa  -7.545e+02  3.697e+02  -2.041 0.041442 *
QualityIndex:BsmtQual.TA  -7.714e+02  2.096e+02  -3.681 0.000242 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28070 on 1380 degrees of freedom
Multiple R-squared:  0.8366,    Adjusted R-squared:  0.8353
F-statistic: 642.1 on 11 and 1380 DF,  p-value: < 2.2e-16
```

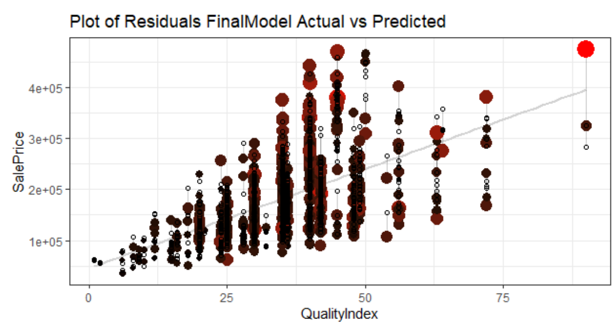
```
> anova (final.2)
Analysis of Variance Table

Response: SalePrice
                            Df     Sum Sq    Mean Sq  F value
QualityIndex                 1 1.8456e+12 1.8456e+12 2342.3297
TotalSqftCalc                1 2.8039e+12 2.8039e+12 3558.4472
BsmtQual.Ex                  1 4.4001e+11 4.4001e+11  558.4209
BsmtQual.Fa                  1 1.3821e+10 1.3821e+10   17.5404
BsmtQual.TA                  1 3.6175e+11 3.6175e+11  459.1037
TotalSqftCalc:BsmtQual.Ex    1 3.0602e+10 3.0602e+10   38.8381
TotalSqftCalc:BsmtQual.Fa    1 1.4780e+09 1.4780e+09    1.8758
TotalSqftCalc:BsmtQual.TA    1 2.6068e+09 2.6068e+09    3.3083
QualityIndex:BsmtQual.Ex     1 5.4868e+10 5.4868e+10   69.6344
QualityIndex:BsmtQual.Fa     1 2.1485e+08 2.1485e+08    0.2727
QualityIndex:BsmtQual.TA     1 1.0674e+10 1.0674e+10   13.5463
Residuals                 1380 1.0874e+12 7.8795e+08
                             Pr(>F)
QualityIndex              < 2.2e-16 ***
TotalSqftCalc             < 2.2e-16 ***
BsmtQual.Ex               < 2.2e-16 ***
BsmtQual.Fa               2.991e-05 ***
BsmtQual.TA               < 2.2e-16 ***
TotalSqftCalc:BsmtQual.Ex 6.105e-10 ***
TotalSqftCalc:BsmtQual.Fa 0.1710347
TotalSqftCalc:BsmtQual.TA 0.0691470 .
QualityIndex:BsmtQual.Ex  < 2.2e-16 ***
QualityIndex:BsmtQual.Fa  0.6016270
QualityIndex:BsmtQual.TA  0.0002417 ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $Y= 14380+2343\beta_1-50.49\beta_2-172200\beta_3-5784\beta_4-7326\beta_5+12.71\beta_6-9.34\beta_7-1.80\beta_8+48523\beta_9-754.5\beta_{10}-771.4\beta_{11}$
- If all held constant using BsmTQuality is Great as the basis of interpretation
- $H_0 : \beta_0 = \beta_1 \ldots \beta_{11}= 0$
- $H_a : \beta_j \neq 0$, for at least one value of j (for j in 0, 1…11)
- We reject null hypothesis, based on $F(11, 1380) = 642.1$, p-value: < 2.2e-16. The model is statistically significant in explaining 83.66% of the variance of response variable SalePrice.
- Actual versus predicted SalePrice side by side view presented by TotalSqftCalc and QualityIndex shows visually below.
- Residuals are normally distributed based on histogram
- QQ plot shows normality except around the tails due to extreme outliers. (1098, 1832, 1200)
- Residual and Fitted graph show a slight parabola shape, so we can say the model does not meet the homoscedasticity assumption since the residuals are not equally spread around the y = 0 line. Linear model assumption is there since the redline through our scatterplot is fairly straight.
- Scale location graph is supporting the evaluation of homoscedasticity. We see the red line is sloping slightly curved, but the data points are randomly spread out.
- Cook's distance shows other than few extreme values, we see most within the range.




Lastly, let's look at the model final and revised model final.2 in terms of MAE and MSE comparison, and model including the effect of basement quality with TotalSqftCalc and QualityIndex is a better fit than initial final model does not including the effect variables.

| | Train | | Test | |
|---|---|---|---|---|
| | MAE | MSE | MAE | MSE |
| Final model | 21726.02 | 853314781 | 21671.41 | 859201358 |
| Final model + Effect | 20755.06 | 781156538 | 20408.12 | 763794594 |

7) *For reflection / conclusions:*

I think the biggest challenge and require most effort is the pre-work, which is determining the various variables that should be included in the modeling exercise. Naturally we want to include as much as we can out of 89 total variables. But interestingly, as we strip away the variables one at a time, we realized, much simpler model was lot easier to interpret. When selecting variables, using junk model, we learned to not include variable if they are being derived from one another. VIF score needs to be included in addition to purely relying on coefficients.

In this analysis, I also spend significant amount of time in the back end, since I am keeping a categorical variable and need to continue my analysis in understanding the effects among the 2 numerical variables since we have the ANOVA model on hand. At the end, it's a great learning process by understanding the step by step of analyzing ANOVA model from last week and applying it into this final model build.

Out of curiosity, I re-evaluated the model by dropping BsmtQual variable completely, and left with 2 numerical variables. With that simpler model, we can explain the variance of SalePrice by 69.89%, and forego the lengthy ANOVA model evaluation.

If we want a simpler model, and easy to interpret then that can be an option. However, we are facing following issues.

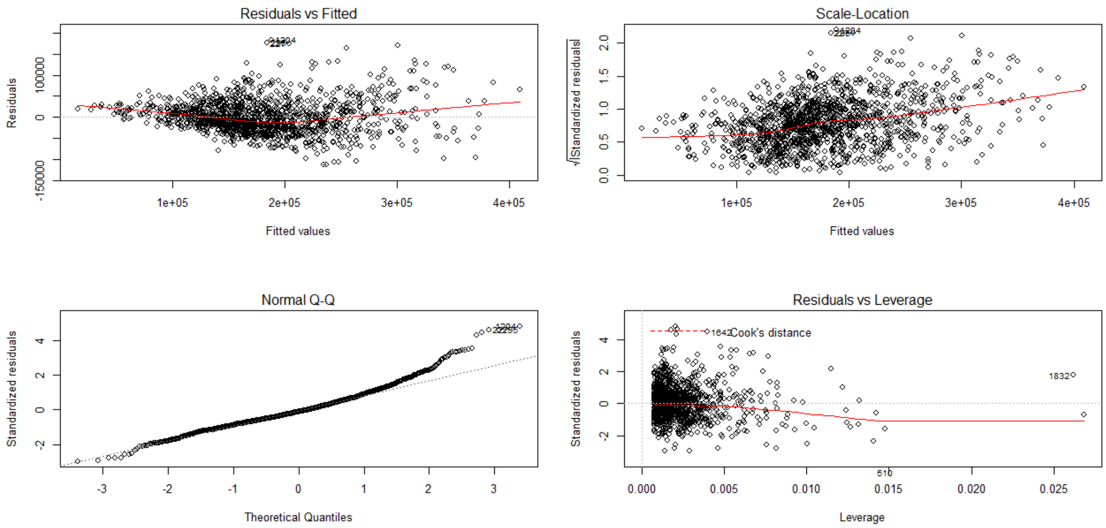
1. The diagnosis suggests the simple model violates a lot of the assumptions.
2. Mean absolute error is much higher than our revised final model.
3. Prediction grade shows we no longer can claim underwriting quality. Since the model is not accurate 10% within 10% more than 50% of the time.

So simpler is better is valid to certain extent, as a modeler we have to balance and interpretation and usability and accuracy of the model that we are creating and suggesting to the business.

```
Call:
lm(formula = SalePrice ~ QualityIndex + TotalSqftCalc, data = train.clean1)

Residuals:
    Min      1Q  Median      3Q     Max
-112762  -23935   -4062   20685  183187

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -31069.207   4228.842  -7.347 3.44e-13 ***
QualityIndex    2110.172    115.844  18.216  < 2e-16 ***
TotalSqftCalc     69.325      1.572  44.091  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37980 on 1389 degrees of freedom
Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6984
F-statistic:  1612 on 2 and 1389 DF,  p-value: < 2.2e-16
```

Model Comparison

| | Train | | Test | |
|---|---|---|---|---|
| | MAE | MSE | MAE | MSE |
| Final model | 21726.02 | 853314781 | 21671.41 | 859201358 |
| Final model + Effect | 20755.06 | 781156538 | 20408.12 | 763794594 |
| R12(TotalSqftCalc+QualityIndex) | 28600.16 | 1439218924 | 28144.29 | 1297658840 |




**Prediction Grade**

| r12.testPredictionGrade | | | |
|---|---|---|---|
| Grade 1: [0.0.10] | Grade 2: (0.10,0.15] | Grade 3: (0.15,0.25] | Grade 4: (0.25+] |
| 0.3633277 | 0.1629881 | 0.2444822 | 0.229202 |