Michael Venit

MSDS 410

Modeling Assignment 3

1. Below we can see all the categorical variables in the dataset:

```
 [1] "Zoning"        "Street"        "Alley"         "LotShape"      "LandContour"   "Utilities"    
"LotConfig"    
 [8] "LandSlope"     "Neighborhood"  "Condition1"    "Condition2"    "BldgType"      "HouseStyle"   
"RoofStyle"    
[15] "RoofMat"       "Exterior1"     "Exterior2"     "MasVnrType"    "ExterQual"     "ExterCond"    
"Foundation"   
[22] "BsmtQual"      "BsmtCond"      "BsmtExposure"  "BsmtFinType1"  "BsmtFinType2"  "Heating"      
"HeatingQC"    
[29] "CentralAir"    "Electrical"    "KitchenQual"   "Functional"    "FireplaceQu"   "GarageType"   
"GarageFinish" 
[36] "GarageQual"    "GarageCond"    "PavedDrive"    "PoolQC"        "Fence"         "MiscFeature"  
"SaleType"     
[43] "SaleCondition"
```

At first glance, the categorical variables that seem most interesting are Neighborhood and Condition1. Intuitively, these are that should be looked at when examining a property.

Condition1 pertains to the proximity of the property to various city conditions described below:

Condition1	**Definition**
Norm	Normal
Feedr	Adjacent to feeder street
PosN	Near positive off-site feature–park, greenbelt, etc.
RRNe	Within 200’ of East-West Railroad
RRAe	Adjacent to East-West Railroad
Artery	Adjacent to arterial street
PosA	Adjacent to positive off-site feature
RRAn	Adjacent to North-South Railroad
RRNn	Within 200’ of North-South Railroad

Summary statistics for the Neighborhood feature can also be seen in the following table:

Neighborhood	**MeanSP**	**MedSP**	**SdSP**
Blmngtn	159895.0	159895	NA
BrkSide	126740.4	127750	36626.34
ClearCr	218400.9	225000	49440.80
CollgCr	199779.2	200500	46076.73
Crawfor	199021.4	196500	58024.03
Edwards	132956.2	125000	50769.51
Gilbert	189209.6	184050	28546.83
IDOTRR	121108.1	120500	31454.23
Mitchel	166527.1	156450	41942.28
NAmes	146903.7	142000	30603.09
NoRidge	319616.0	301750	73717.55
NridgHt	345267.9	326000	84852.62
NWAmes	194384.1	185000	35990.01
OldTown	128551.8	122000	47275.61
Sawyer	137326.1	135000	22903.23
SawyerW	190508.2	184900	47471.48
Somerst	248517.9	245000	42214.65
StoneBr	348963.1	349265	92916.67
SWISU	132983.8	135750	31331.94
Timber	241995.2	214900	72156.23

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           159895      45625   3.505 0.000468 ***
NeighborhoodBrkSide   -33155      45862  -0.723 0.469811
NeighborhoodClearCr    58506      46237   1.265 0.205898
NeighborhoodCollgCr    39884      45732   0.872 0.383241
NeighborhoodCrawfor    39126      45916   0.852 0.394247
NeighborhoodEdwards   -26939      45801  -0.588 0.556487
NeighborhoodGilbert    29315      45803   0.640 0.522233
NeighborhoodIDOTRR    -38787      46079  -0.842 0.400028
NeighborhoodMitchel     6632      45899   0.144 0.885125
NeighborhoodNAmes     -12991      45688  -0.284 0.776173
NeighborhoodNoRidge   159721      45969   3.475 0.000523 ***
NeighborhoodNridgHt   185373      45964   4.033 5.72e-05 ***
NeighborhoodNWAmes     34489      45826   0.753 0.451776
NeighborhoodOldTown   -31343      45754  -0.685 0.493403
NeighborhoodSawyer    -22569      45813  -0.493 0.622327
NeighborhoodSawyerW    30613      45880   0.667 0.504697
NeighborhoodSomerst    88623      45964   1.928 0.053987 .
NeighborhoodStoneBr   189068      47347   3.993 6.76e-05 ***
NeighborhoodSWISU     -26911      46291  -0.581 0.561071
NeighborhoodTimber     82100      46088   1.781 0.075004 .
NeighborhoodVeenker    92596      46947   1.972 0.048711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45620 on 1966 degrees of freedom
Multiple R-squared:  0.6021,	Adjusted R-squared:  0.598
F-statistic: 148.7 on 20 and 1966 DF,  p-value: < 2.2e-16
```

When observing the summary of the regression model SalePrice ~ Neighborhood (as seen above), we can review the coefficients as the mean difference from Blmngtn. There appear to be several neighborhoods with very big mean differences, while the adjusted R-squared also indicates that Neighborhood accounts for a high amount of variance in SalePrice. Meanwhile, a summary table for Condition1 can be found below, in which there is a wide range of mean SalePrice amongst the conditions.

Condition1	MeanSP	MedSP	SdSP
Artery	129209.4	120000	57221.35
Feedr	139481.8	137500	40744.30
Norm	183153.9	165500	72098.52
PosA	235075.0	191000	102494.90
PosN	226256.5	206500	88658.65
RRAe	141645.0	142950	20788.52
RRAn	172029.3	170250	45028.96
RRNe	150337.5	156500	48571.65
RRNn	209208.3	217000	73272.51

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       129209       8778  14.721  < 2e-16 ***
Condition1Feedr    10272      11021   0.932  0.35142    
Condition1Norm     53944       8941   6.033 1.91e-09 ***
Condition1PosA    105866      18734   5.651 1.83e-08 ***
Condition1PosN     97047      14902   6.512 9.35e-11 ***
Condition1RRAe     12436      17988   0.691  0.48945    
Condition1RRAn     42820      15203   2.817  0.00490 ** 
Condition1RRNe     21128      36190   0.584  0.55942    
Condition1RRNn     79999      29981   2.668  0.00769 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70220 on 1978 degrees of freedom
Multiple R-squared:  0.05166,	Adjusted R-squared:  0.04783 
F-statistic: 13.47 on 8 and 1978 DF,  p-value: < 2.2e-16
```

The regression summary confirms our statement above, in that there are big differences in the means amongst the levels of Condition1. However, the adjusted R-squared is near zero which indicates that the Condition1 variable does not account for very much, if any, variance in SalePrice. Hence, this variable will not be used in further analysis.

2. The number of observations and percentage of total observations in the partitioned training and test data frames can be seen in the following table:

DataFrame	ObsCounts	PercentOfObs
train_df	1398	0.7035732
test_df	589	0.2964268

3. The features that have been chosen will be included in the pool of candidate predictors, which can be seen below:

TrainDfVariables
SalePrice
YrSold
TotalSqftCalc
LotFrontage
LotArea
QualityIndex
TotalBsmtSF
FullBath
MasVnrArea
YearRemodel
BedroomAbvGr
GarageCars
BsmtFinishRatio
WoodDeckSF
GarageArea
PoolArea
TotRmsAbvGrd

At this point, it was realized that there were roughly 300 unique observations in the training data that had missing values for some of the selected features. Since this constituted greater than 20% of my training data, imputation was enacted on the LotFrontage, MasVnrArea and BsmtFinishRatio variables. The imputed values that were used came to be the respective median values for each feature. As it pertains to model identification, 3 models were created: forward, backwards and stepwise. The final model summary for forward.lm can be seen below along with the VIF values:

```
lm(formula = SalePrice ~ TotalSqftCalc + BsmtFinishRatio + GarageCars +
    YearRemodel + MasVnrArea + QualityIndex + TotalBsmtSF + LotArea +
    BedroomAbvGr + TotRmsAbvGrd + WoodDeckSF + FullBath + LotFrontage +
    GarageArea, data = train_clean_df)

Residuals:
    Min      1Q  Median      3Q     Max
-100144  -14866    -432   14770  170491

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.374e+05  8.311e+04 -11.279  < 2e-16 ***
TotalSqftCalc    5.415e+01  2.731e+00  19.832  < 2e-16 ***
BsmtFinishRatio -3.575e+04  3.441e+03 -10.390  < 2e-16 ***
GarageCars       1.114e+04  2.225e+03   5.006 6.26e-07 ***
YearRemodel      4.626e+02  4.287e+01  10.792  < 2e-16 ***
MasVnrArea       5.689e+01  5.178e+00  10.987  < 2e-16 ***
QualityIndex     1.044e+03  9.335e+01  11.182  < 2e-16 ***
TotalBsmtSF      2.018e+01  2.692e+00   7.496 1.17e-13 ***
LotArea          5.708e-01  1.099e-01   5.193 2.38e-07 ***
BedroomAbvGr    -9.212e+03  1.382e+03  -6.664 3.83e-11 ***
TotRmsAbvGrd     3.114e+03  9.694e+02   3.212 0.001348 **
WoodDeckSF       1.914e+01  5.743e+00   3.332 0.000885 ***
FullBath         6.272e+03  1.985e+03   3.159 0.001618 **
LotFrontage      1.124e+02  4.479e+01   2.509 0.012221 *
GarageArea       1.868e+01  7.696e+00   2.427 0.015357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26800 on 1383 degrees of freedom
Multiple R-squared:  0.8713,	Adjusted R-squared:   0.87
F-statistic: 668.6 on 14 and 1383 DF,  p-value: < 2.2e-16
```

	forward_vif
TotalSqftCalc	7.705857
BsmtFinishRatio	2.771697
GarageCars	5.119003
YearRemodel	1.535678
MasVnrArea	1.475865
QualityIndex	1.374242
TotalBsmtSF	2.228758
LotArea	1.204116
BedroomAbvGr	1.829330
TotRmsAbvGrd	3.845721
WoodDeckSF	1.166551
FullBath	2.269036
LotFrontage	1.300421
GarageArea	4.798364

Following the forward.lm model, the generation of the backwards.lm model and VIF values were the next step in the model identification process.

```
lm(formula = SalePrice ~ TotalSqftCalc + LotFrontage + LotArea +
    QualityIndex + TotalBsmtSF + FullBath + MasVnrArea + YearRemodel +
    BedroomAbvGr + GarageCars + BsmtFinishRatio + WoodDeckSF +
    GarageArea + TotRmsAbvGrd, data = train_clean_df)

Residuals:
    Min      1Q  Median      3Q     Max
-100144  -14866    -432   14770  170491

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.374e+05  8.311e+04 -11.279  < 2e-16 ***
TotalSqftCalc    5.415e+01  2.731e+00  19.832  < 2e-16 ***
LotFrontage      1.124e+02  4.479e+01   2.509 0.012221 *
LotArea          5.708e-01  1.099e-01   5.193 2.38e-07 ***
QualityIndex     1.044e+03  9.335e+01  11.182  < 2e-16 ***
TotalBsmtSF      2.018e+01  2.692e+00   7.496 1.17e-13 ***
FullBath         6.272e+03  1.985e+03   3.159 0.001618 **
MasVnrArea       5.689e+01  5.178e+00  10.987  < 2e-16 ***
YearRemodel      4.626e+02  4.287e+01  10.792  < 2e-16 ***
BedroomAbvGr    -9.212e+03  1.382e+03  -6.664 3.83e-11 ***
GarageCars       1.114e+04  2.225e+03   5.006 6.26e-07 ***
BsmtFinishRatio -3.575e+04  3.441e+03 -10.390  < 2e-16 ***
WoodDeckSF       1.914e+01  5.743e+00   3.332 0.000885 ***
GarageArea       1.868e+01  7.696e+00   2.427 0.015357 *
TotRmsAbvGrd     3.114e+03  9.694e+02   3.212 0.001348 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26800 on 1383 degrees of freedom
Multiple R-squared:  0.8713,	Adjusted R-squared:   0.87
F-statistic: 668.6 on 14 and 1383 DF,  p-value: < 2.2e-16
```

	back_vif
TotalSqftCalc	7.705857
LotFrontage	1.300421
LotArea	1.204116
QualityIndex	1.374242
TotalBsmtSF	2.228758
FullBath	2.269036
MasVnrArea	1.475865
YearRemodel	1.535678
BedroomAbvGr	1.829330
GarageCars	5.119003
BsmtFinishRatio	2.771697
WoodDeckSF	1.166551
GarageArea	4.798364
TotRmsAbvGrd	3.845721

The next step in the model identification process is the creation of the stepwise model, which can be seen below with corresponding VIF values. There are no values over 10, however there are a couple variables between 5 and 8 which could be indicative of collinearity.

```
lm(formula = SalePrice ~ TotalSqftCalc + BsmtFinishRatio + GarageCars +
    YearRemodel + MasVnrArea + QualityIndex + TotalBsmtSF + LotArea +
    BedroomAbvGr + TotRmsAbvGrd + WoodDeckSF + FullBath + LotFrontage +
    GarageArea, data = train_clean_df)

Residuals:
    Min      1Q  Median      3Q     Max
-100144  -14866    -432   14770  170491

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.374e+05  8.311e+04 -11.279  < 2e-16 ***
TotalSqftCalc    5.415e+01  2.731e+00  19.832  < 2e-16 ***
BsmtFinishRatio -3.575e+04  3.441e+03 -10.390  < 2e-16 ***
GarageCars       1.114e+04  2.225e+03   5.006 6.26e-07 ***
YearRemodel      4.626e+02  4.287e+01  10.792  < 2e-16 ***
MasVnrArea       5.689e+01  5.178e+00  10.987  < 2e-16 ***
QualityIndex     1.044e+03  9.335e+01  11.182  < 2e-16 ***
TotalBsmtSF      2.018e+01  2.692e+00   7.496 1.17e-13 ***
LotArea          5.708e-01  1.099e-01   5.193 2.38e-07 ***
BedroomAbvGr    -9.212e+03  1.382e+03  -6.664 3.83e-11 ***
TotRmsAbvGrd     3.114e+03  9.694e+02   3.212 0.001348 **
WoodDeckSF       1.914e+01  5.743e+00   3.332 0.000885 ***
FullBath         6.272e+03  1.985e+03   3.159 0.001618 **
LotFrontage      1.124e+02  4.479e+01   2.509 0.012221 *
GarageArea       1.868e+01  7.696e+00   2.427 0.015357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26800 on 1383 degrees of freedom
Multiple R-squared:  0.8713,    Adjusted R-squared:   0.87
F-statistic: 668.6 on 14 and 1383 DF,  p-value: < 2.2e-16
```

	stepwise_vif
TotalSqftCalc	7.705857
BsmtFinishRatio	2.771697
GarageCars	5.119003
YearRemodel	1.535678
MasVnrArea	1.475865
QualityIndex	1.374242
TotalBsmtSF	2.228758
LotArea	1.204116
BedroomAbvGr	1.829330
TotRmsAbvGrd	3.845721
WoodDeckSF	1.166551
FullBath	2.269036
LotFrontage	1.300421
GarageArea	4.798364

This next model is referred to as junk because the independent variables will be highly correlated due to the fact that the QualityIndex feature is made up of the other two Quality variables. There will be a high level of collinearity between the three Quality variables. We can see from the VIF values below that there is high collinearity between the quality variables.

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

Residuals:
    Min      1Q  Median      3Q     Max
-129293  -16495   -1434   14634  187296

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.991e+05  1.738e+04 -11.455  < 2e-16 ***
OverallQual    4.407e+04  2.942e+03  14.978  < 2e-16 ***
OverallCond    1.952e+04  3.126e+03   6.245 5.63e-10 ***
QualityIndex  -3.518e+03  5.333e+02  -6.597 5.96e-11 ***
GrLivArea      2.552e+01  2.731e+00   9.345  < 2e-16 ***
TotalSqftCalc  4.260e+01  1.799e+00  23.684  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29110 on 1392 degrees of freedom
Multiple R-squared:  0.8471,    Adjusted R-squared:  0.8465
F-statistic:  1542 on 5 and 1392 DF,  p-value: < 2.2e-16
```

	junk_vif
OverallQual	24.980307
OverallCond	20.510484
QualityIndex	38.004574
GrLivArea	3.112674
TotalSqftCalc	2.833495

As it turns out, the three methods all selected the same model. Following the creation of these models, a crucial step in the identification process is model comparison. A table summarizing each of the model’s adjusted R-squared, AIC, BIC, MSE, RMSE and MAE values can be seen below:

Model	AdjRSquared	AIC	BIC	MSE	RMSE	MAE
Forward	0.8699651	32492.54	32576.43	710439352	26654.07	19534.48
Backward	0.8699651	32492.54	32576.43	710439352	26654.07	19534.48
Stepwise	0.8699651	32492.54	32576.43	710439352	26654.07	19534.48
Junk	0.8465204	32715.35	32752.05	843985070	29051.42	21208.36

A ranking was not added to the data frame due to the fact that each of the 3 model selection methods chose the same model. Therefore, the metrics were the same. The junk model performed the worst in each metric.

4. Prediction accuracy is now of interest as it pertains to each of the models created above. MAE and MSE were computed for the respective models on the test data, which can be seen in the following table:

Model	TestMSE	TestMAE
Forward	722289748	19118.70
Backward	722289748	19118.70
Stepwise	722289748	19118.70
Junk	908959192	21450.31

The three models chosen by automatic variable selection all performed the same (since they are the same model). The junk model performed the worst on both metrics. The MSE is higher on the test data, and the MAE is lower on the test data. It would seem that when the MSE is lower on the training data, the model is over-fitting the training data.

5. The models have been validated in a statistical sense, but we would like to generate prediction grades for operational validation. Prediction grades, on the training and test data, for the forward model and junk model created can be found below (forward, backwards and step models are same):

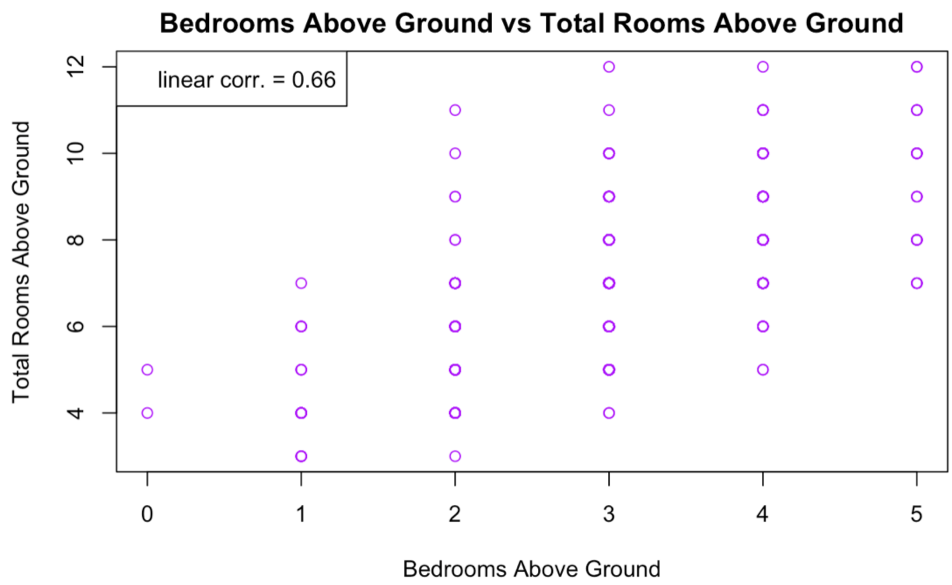
```
forward_PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.5565093            0.1716738            0.1702432            0.1015737

junk_PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.5200286            0.1816881            0.1759657            0.1223176

forward_testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
         0.55857385           0.16468591           0.18336163           0.09337861

junk_testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.5348048            0.1595925            0.1782683            0.1273345
```

I only built the prediction grade for the forward.lm model since all the models were the same. Over 70% of the predictions on the test set were within 15% of the actual observation. Additionally, roughly 56% of the predictions on the test set were within 50%, which would fall under the "underwriting quality" category.

6. Now that we have the “best” model for the purpose of this assignment, it is important to revisit and clean-up the model. The stepwise model was chosen, determined by the stepwise automated variable selection. I chose this model somewhat arbitrarily due to the fact that all three methods identified the same model. I will start by looking at the regression coefficients:

```
lm(formula = SalePrice ~ TotalSqftCalc + BsmtFinishRatio + GarageCars +
    YearRemodel + MasVnrArea + QualityIndex + TotalBsmtSF + LotArea +
    BedroomAbvGr + TotRmsAbvGrd + WoodDeckSF + FullBath + LotFrontage +
    GarageArea, data = train_clean_df)

Residuals:
    Min      1Q  Median      3Q     Max
-100144  -14866    -432   14770  170491

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.374e+05  8.311e+04 -11.279  < 2e-16 ***
TotalSqftCalc    5.415e+01  2.731e+00  19.832  < 2e-16 ***
BsmtFinishRatio -3.575e+04  3.441e+03 -10.390  < 2e-16 ***
GarageCars       1.114e+04  2.225e+03   5.006 6.26e-07 ***
YearRemodel      4.626e+02  4.287e+01  10.792  < 2e-16 ***
MasVnrArea       5.689e+01  5.178e+00  10.987  < 2e-16 ***
QualityIndex     1.044e+03  9.335e+01  11.182  < 2e-16 ***
TotalBsmtSF      2.018e+01  2.692e+00   7.496 1.17e-13 ***
LotArea          5.708e-01  1.099e-01   5.193 2.38e-07 ***
BedroomAbvGr    -9.212e+03  1.382e+03  -6.664 3.83e-11 ***
TotRmsAbvGrd     3.114e+03  9.694e+02   3.212 0.001348 **
WoodDeckSF       1.914e+01  5.743e+00   3.332 0.000885 ***
FullBath         6.272e+03  1.985e+03   3.159 0.001618 **
LotFrontage      1.124e+02  4.479e+01   2.509 0.012221 *
GarageArea       1.868e+01  7.696e+00   2.427 0.015357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26800 on 1383 degrees of freedom
Multiple R-squared:  0.8713,	Adjusted R-squared:   0.87
F-statistic: 668.6 on 14 and 1383 DF,  p-value: < 2.2e-16
```

The intercept term in this model does not hold any practical value as it means that the SalePrice would be equal to a negative value if all the variables were equal to zero. At first glance there are some variables that don't make much sense, as BedroomAbvGr has a negative coefficient but TotRmsAbvGrd has a positive coefficient. This means that as the number of bedrooms above ground increases, the Sale Price decreases. The corresponding TotRmsAbvGrd variable would indicate that as total rooms above ground increases, the Sale Price would increase. These two variables should be heading in the same direction. There is possibly some collinearity going on here, hence it is worth investigating the relationship between the two variables.



There's obviously a positive relationship between these two variables. After observing this relationship, dropping BedroomAbvGr variable and seeing if it dramatically affects the R-squared value seems like an appropriate next step. The model summary, excluding the BedroomAbvGr feature, can be seen below:

```
lm(formula = SalePrice ~ TotalSqftCalc + BsmtFinishRatio + GarageCars +
    YearRemodel + MasVnrArea + QualityIndex + TotalBsmtSF + LotArea +
    TotRmsAbvGrd + WoodDeckSF + FullBath + LotFrontage + GarageArea,
    data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-94471 -14475  -1092  14199 186602

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.809e+05  8.415e+04 -11.657  < 2e-16 ***
TotalSqftCalc    5.366e+01  2.772e+00  19.358  < 2e-16 ***
BsmtFinishRatio -3.564e+04  3.495e+03 -10.197  < 2e-16 ***
GarageCars       1.188e+04  2.257e+03   5.262 1.65e-07 ***
YearRemodel      4.799e+02  4.345e+01  11.045  < 2e-16 ***
MasVnrArea       5.716e+01  5.258e+00  10.870  < 2e-16 ***
QualityIndex     1.062e+03  9.476e+01  11.211  < 2e-16 ***
TotalBsmtSF      2.144e+01  2.728e+00   7.862 7.58e-15 ***
LotArea          6.000e-01  1.115e-01   5.379 8.78e-08 ***
TotRmsAbvGrd     1.656e+02  8.760e+02   0.189 0.850072
WoodDeckSF       2.046e+01  5.829e+00   3.511 0.000461 ***
FullBath         5.054e+03  2.008e+03   2.517 0.011938 *
LotFrontage      1.157e+02  4.548e+01   2.543 0.011088 *
GarageArea       1.872e+01  7.816e+00   2.396 0.016721 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27220 on 1384 degrees of freedom
Multiple R-squared:  0.8671,	Adjusted R-squared:  0.8659
F-statistic: 694.8 on 13 and 1384 DF,  p-value: < 2.2e-16
```

From the reduced model, removing the BedroomAbvGr variable does not impact the adjusted R-squared much, hence it can be safely removed from the previous model to avoid a difficult to interpret coefficient. The BsmtFinishRatio feature also has a very large negative coefficient indicating that as the ratio of finished basement space to total basement space increases one unit, the Sale price will decrease by ~$35k (when all other variables held constant). This is another variable that doesn't make sense as one would expect the coefficient to be a positive value. Intuitively, as the usable space in the basement increases proportional to the total basement space, I would expect that to improve the value of a home. Again, this is another feature that can be explored for removal from the reduced model to view if it has an impact on the adjusted R-squared value. This can be seen from the following model summary below:

```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF + LotArea + TotRmsAbvGrd +
    WoodDeckSF + FullBath + LotFrontage + GarageArea, data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-97244 -15570   -848  14972 209580

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.553e+05  8.718e+04 -10.958  < 2e-16 ***
TotalSqftCalc  3.186e+01  1.829e+00  17.421  < 2e-16 ***
GarageCars     1.201e+04  2.340e+03   5.135 3.22e-07 ***
YearRemodel    4.518e+02  4.495e+01  10.052  < 2e-16 ***
MasVnrArea     6.272e+01  5.421e+00  11.571  < 2e-16 ***
QualityIndex   1.214e+03  9.700e+01  12.514  < 2e-16 ***
TotalBsmtSF    3.280e+01  2.581e+00  12.710  < 2e-16 ***
LotArea        7.004e-01  1.152e-01   6.082 1.53e-09 ***
TotRmsAbvGrd   5.154e+03  7.533e+02   6.842 1.17e-11 ***
WoodDeckSF     1.876e+01  6.039e+00   3.106  0.00194 **
FullBath       1.046e+04  2.007e+03   5.211 2.16e-07 ***
LotFrontage    8.747e+01  4.706e+01   1.859  0.06324 .
GarageArea     1.892e+01  8.101e+00   2.336  0.01963 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28210 on 1385 degrees of freedom
Multiple R-squared:  0.8572,    Adjusted R-squared:  0.8559
F-statistic: 692.5 on 12 and 1385 DF,  p-value: < 2.2e-16
```

Compared the first reduced model, the adjusted R-squared only decreased by 1% (0.8659 down to 0.8559), designating that the BsmtFinishRatio variable was only accounting for an additional 1% of variance in the Sale Price. I will keep this variable removed from the final model due to its low account of variance and its difficult coefficient interpretation. While looking at the model summary, one coefficients is really small, as the LotArea variable has a coefficient of 0.7004. This means that as the LotArea increases one unit (or one Square foot), the Sale price only increases 7 cents. This interpretation seems quite useless in our current model as there would have to be a dramatically bigger/smaller lot for it to have any noticeable impact. This is another feature that can be explored for removal, by observing the change in the adjusted R-squared value.

```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF + TotRmsAbvGrd +
    WoodDeckSF + FullBath + LotFrontage + GarageArea, data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-95829 -16187  -1037  14765 205403

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.274e+05  8.818e+04 -10.517  < 2e-16 ***
TotalSqftCalc  3.357e+01  1.831e+00  18.338  < 2e-16 ***
GarageCars     1.223e+04  2.369e+03   5.162 2.80e-07 ***
YearRemodel    4.371e+02  4.546e+01   9.615  < 2e-16 ***
MasVnrArea     5.986e+01  5.470e+00  10.943  < 2e-16 ***
QualityIndex   1.203e+03  9.824e+01  12.244  < 2e-16 ***
TotalBsmtSF    3.319e+01  2.613e+00  12.703  < 2e-16 ***
TotRmsAbvGrd   5.207e+03  7.629e+02   6.825 1.31e-11 ***
WoodDeckSF     2.078e+01  6.108e+00   3.403 0.000686 ***
FullBath       1.055e+04  2.033e+03   5.188 2.44e-07 ***
LotFrontage    1.530e+02  4.640e+01   3.298 0.000998 ***
GarageArea     1.862e+01  8.205e+00   2.270 0.023381 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28570 on 1386 degrees of freedom
Multiple R-squared:  0.8533,    Adjusted R-squared:  0.8522
F-statistic: 733.1 on 11 and 1386 DF,  p-value: < 2.2e-16
```

The removal of the LotArea coefficient only decreased the adjusted R-squared by 0.0037, signifying that LotArea almost has zero predictive power. This featrure can safely be removed without any effect on the final model. At this point, 3 variables have been removed with only an overall drop in adjusted R-squared of 0.0178. We are on our way to a much more parsimonious model that will be much easier to explain. The next variable that will be explored for removal will GarageArea and the resulting model can be seen below:

```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF + TotRmsAbvGrd +
    WoodDeckSF + FullBath + LotFrontage, data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-93549 -16061  -1022  14808 203719

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.271e+05  8.831e+04 -10.498  < 2e-16 ***
TotalSqftCalc  3.401e+01  1.823e+00  18.654  < 2e-16 ***
GarageCars     1.660e+04  1.382e+03  12.010  < 2e-16 ***
YearRemodel    4.372e+02  4.553e+01   9.603  < 2e-16 ***
MasVnrArea     6.058e+01  5.469e+00  11.077  < 2e-16 ***
QualityIndex   1.206e+03  9.838e+01  12.258  < 2e-16 ***
TotalBsmtSF    3.342e+01  2.615e+00  12.780  < 2e-16 ***
TotRmsAbvGrd   5.069e+03  7.616e+02   6.655 4.06e-11 ***
WoodDeckSF     2.082e+01  6.117e+00   3.403 0.000686 ***
FullBath       1.041e+04  2.035e+03   5.116 3.56e-07 ***
LotFrontage    1.584e+02  4.641e+01   3.413 0.000662 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28620 on 1387 degrees of freedom
Multiple R-squared:  0.8528,    Adjusted R-squared:  0.8517
F-statistic: 803.5 on 10 and 1387 DF,  p-value: < 2.2e-16
```

From the resulting model summary, the GarageArea variable only accounted for a small amount of variance in SalePrice (0.8522 down to 0.8517), hence it would be safe for removal from the final model. Next, I will remove the LotFrontage which represents the exposure of the property to the street. It possesses the lowest t-value out of the remaining variables hence it is work exploring its removal from the reduced model.

```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF + TotRmsAbvGrd +
    WoodDeckSF + FullBath, data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-91628 -16292  -1393  15517 204678

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.201e+05  8.863e+04 -10.382  < 2e-16 ***
TotalSqftCalc  3.460e+01  1.822e+00  18.994  < 2e-16 ***
GarageCars     1.724e+04  1.375e+03  12.536  < 2e-16 ***
YearRemodel    4.379e+02  4.570e+01   9.581  < 2e-16 ***
MasVnrArea     6.314e+01  5.438e+00  11.612  < 2e-16 ***
QualityIndex   1.180e+03  9.847e+01  11.988  < 2e-16 ***
TotalBsmtSF    3.386e+01  2.622e+00  12.915  < 2e-16 ***
TotRmsAbvGrd   5.267e+03  7.623e+02   6.908 7.44e-12 ***
WoodDeckSF     2.032e+01  6.139e+00   3.311 0.000954 ***
FullBath       1.027e+04  2.043e+03   5.030 5.54e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28730 on 1388 degrees of freedom
Multiple R-squared:  0.8516,	Adjusted R-squared:  0.8506
F-statistic: 884.7 on 9 and 1388 DF,  p-value: < 2.2e-16
```

The adjusted squared has only decreased 0.0011 by removing the LotFrontage feature, which indicates that this variable provides little to no predictive value to Sale Price. Hence it will be removed from the final model. Thus far, 5 variables have been removed from the intitial model and the adjusted R-squared has only gone down 0.0194, which seems to be an acceptable drop in order to have a more succinct and understandable model. WoodDeckSF also has a low t-value, hence its removal will be explored next.

```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF + TotRmsAbvGrd +
    FullBath, data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-93837 -16260  -1425  14923 202135

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.500e+05  8.848e+04 -10.736  < 2e-16 ***
TotalSqftCalc  3.572e+01  1.796e+00  19.884  < 2e-16 ***
GarageCars     1.748e+04  1.378e+03  12.685  < 2e-16 ***
YearRemodel    4.529e+02  4.564e+01   9.923  < 2e-16 ***
MasVnrArea     6.300e+01  5.457e+00  11.545  < 2e-16 ***
QualityIndex   1.194e+03  9.873e+01  12.098  < 2e-16 ***
TotalBsmtSF    3.388e+01  2.631e+00  12.876  < 2e-16 ***
TotRmsAbvGrd   5.072e+03  7.628e+02   6.649 4.23e-11 ***
FullBath       1.041e+04  2.049e+03   5.081 4.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28830 on 1389 degrees of freedom
Multiple R-squared:  0.8504,	Adjusted R-squared:  0.8495
F-statistic: 986.8 on 8 and 1389 DF,  p-value: < 2.2e-16
```

The WoodDeckSf variable only contributed a very minor amount to the adjusted R-squared value, so it will be removed from the final model. The next variable to consider is FullBath, again due to the small t-value it possesses.

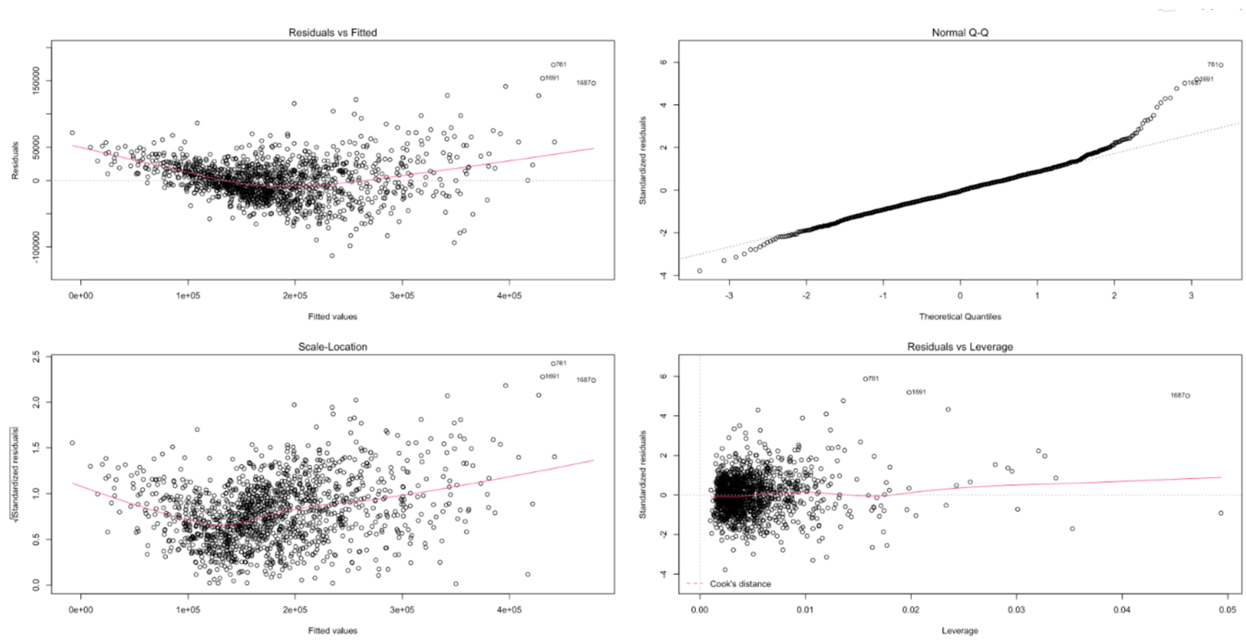
```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF + TotRmsAbvGrd, data = train_clean_df)

Residuals:
   Min     1Q Median     3Q    Max
-99406 -17445  -1014  15913 189968

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.097e+06  8.436e+04 -13.002   <2e-16 ***
TotalSqftCalc  3.678e+01  1.800e+00  20.435   <2e-16 ***
GarageCars     1.905e+04  1.355e+03  14.059   <2e-16 ***
YearRemodel    5.284e+02  4.354e+01  12.136   <2e-16 ***
MasVnrArea     6.304e+01  5.506e+00  11.450   <2e-16 ***
QualityIndex   1.162e+03  9.939e+01  11.687   <2e-16 ***
TotalBsmtSF    3.401e+01  2.654e+00  12.814   <2e-16 ***
TotRmsAbvGrd   6.484e+03  7.166e+02   9.048   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29080 on 1390 degrees of freedom
Multiple R-squared:  0.8476,	Adjusted R-squared:  0.8468
F-statistic:  1104 on 7 and 1390 DF,  p-value: < 2.2e-16
```

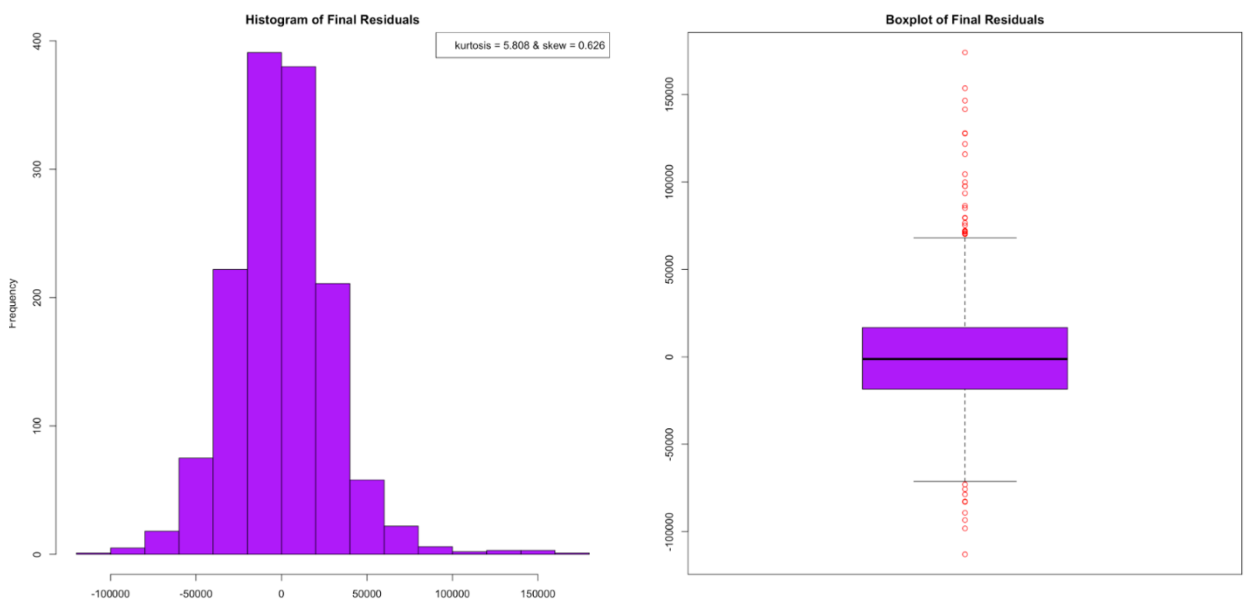
I assumed that FullBath would account for a solid amount of variance in Sale Price, but the adjusted R-squared only decreased by 0.0027, hence it will removed from the final model. From here on out, I will only display the adjusted R-squared values as I remove variables to reduce the clutter of the model summaries.

Removing the TotRmsAbvGrd variable reduced the adjusted R-squared from 0.8468 to 0.8361, resulting in a drop of 0.0107. The adjusted R-squared value has only been reduced by 0.0321 from the original model, despite the removal of 8 features. These 8 variables only accounted for 3.2% of the variance in Sale Price, which is not that significant of an amount. After removing TotBsmtSF, the adjusted R-squared decreased to 0.8245, resulting in a 0.134 drop. This variable will be kept in the final model as of now. The removal of GarageCars and YearRemodel resulted in a reduction of 0.0258 and 0.0183 in adjusted R-squared, leading to the retention of these variables in the final model. If the QualityIndex feature were to be removed, the adjusted R-squared would decrease by 0.017, however, this will be kept in the final model. At this point, the only remaining variable is TotalSqftCalc., though this will not be as it can be derived from the prior tests, and it accounts for the most variance in the Sale Price. The final model is as follows:

```
lm(formula = SalePrice ~ TotalSqftCalc + GarageCars + YearRemodel +
    MasVnrArea + QualityIndex + TotalBsmtSF, data = train_clean_df)

Residuals:
    Min      1Q  Median      3Q     Max
-113010  -18477   -1308   16779  174107

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.137e+06  8.666e+04  -13.12   <2e-16 ***
TotalSqftCalc  4.470e+01  1.618e+00   27.63   <2e-16 ***
GarageCars     2.062e+04  1.382e+03   14.91   <2e-16 ***
YearRemodel    5.614e+02  4.463e+01   12.58   <2e-16 ***
MasVnrArea     6.593e+01  5.654e+00   11.66   <2e-16 ***
QualityIndex   1.259e+03  1.016e+02   12.39   <2e-16 ***
TotalBsmtSF    2.871e+01  2.663e+00   10.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29920 on 1391 degrees of freedom
Multiple R-squared:  0.8386,    Adjusted R-squared:  0.8379
F-statistic:  1205 on 6 and 1391 DF,  p-value: < 2.2e-16
```

Since a final model with reduced features has been generated, the subsequent step would be to preform diagnostic tests on the model. Below we can see the diagnostic graphs from the model. It appears there is possibly a very slight increase in variance with an increase in Y-hat and a histogram of the residuals has been plotted to further investigate this.





The histogram shows that there is a very slight right skew in the residuals, but I do not think this is blatant enough to consider the residuals exhibiting heteroscedasticity. When observing Cook's Distance and leverage, there appears to be slightly over 100 values outside the leverage threshold. The resulting diagnostic tests show that there are no outliers based on Cook's Distance, while there are 103 potential leverage outliers. Overall, this seems to be a pretty good model. It meets the assumptions within reason and unnecessary variables have been eliminated that do not contribute to predicting Sale Price.

7. After working with this data for an extended period of time, the biggest challenges seem to lie in the data wrangling prior to modeling work. This is pretty typical in my experience as far as analytic work goes. To improve predictive accuracy, I would consider going back and including more dummy-coded categorical variables. The trade-off with this route is that it becomes much more work to interpret the model. Generally, I'm a big fan of the motto "simpler is better", as I strive to achieve a level of parsimony. When models become too big or complicated, the interpretation factor also increases. Additionally, if we add a numerous amount of variables that only increase the predictive ability of the model by minuscule amounts, we've unnecessarily complicated our model and we may begin to over-fit it to our data. There is a time and a place for more complicated models and techniques, but I think a lot can be achieved with simpler methods.