Introduction:

In Assignment #5, we will build off of Assignment #3 and develop predictive models for the Ames, Iowa housing data set. This report is divided into six sections. In Section 1, the sample data considered along with drop conditions are explained so that it is clear what is being studied and modeled. In Section 2, framework for predictive modeling is established by splitting the data into training and testing sets. Model identification is performed using various automated variable selection methods and comparing them to a junk model (where predictors are randomly chosen). In Section 3, different fit metrics are compared between the models. In section 4, predictive accuracy on the test data is compared between the models. In section 5, operational validation is assessed to inform business policy development. Diagnostic tests are performed on the best model (in terms of predictive accuracy) as well. In section 6, the final model is tested against OLS assumptions and improved upon (if needed). The report is concluded with a summary of what was learned over the course of these assignments as well as thoughts on how to improve the predictive accuracy of the models.

Section 1: Sample Definition

The Ames Housing dataset is comprised of 82 columns including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. There are also 2 columns (SID and PID) used as observation identifiers. These are not relevant for the purposes of this assignment and were removed. There were also 5 additional variables created. This gives a total of 85 variables used for 2930 residential properties.

It was observed that multiple building types were included in the original data set. It does not make sense to compare one building type to another as each type has a different perception of value in the eyes of a buyer and this is reflected in the different zoning types associated with each building.

The original dataset consists of 5 building types and 8 zoning classifications :

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	2	22	77	2	12	2009	301
2fmCon	0	3	0	0	4	30	25
Duplex	0	0	0	0	4	92	13
Twnhs	0	0	19	0	1	28	53
TwnhsE	0	0	43	0	6	114	70

As the purpose of this assignment is to predict the sale price of a “typical” home, it does not make sense to include buildings falling under the zoning classifications industrial (I), agriculture (A) or commercial (C). Residential zoning types including residential high density (RH), residential low density (RL), residential low density park (RP), residential medium density (RM) and floating village residential will be used for further analysis. For building type, we are looking to analyze houses that represent the typical “house” that a family would buy. Duplexes (Duplx) and townhouses (TwnhsE and TwnhsI) are not good representations and are removed leaving single family detached (1Fam) and two family conversion houses (2FmCon) left for analysis. After these conditions are met, there are 2399 observations remaining. Here is a table showing the distribution of homes in the four zoning classifications by building type.

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	0	0	77	0	12	2004	301
2fmCon	0	0	0	0	0	0	0
Duplex	0	0	0	0	0	0	0
Twnhs	0	0	0	0	0	0	0
TwnhsE	0	0	0	0	0	0	0

It is evident that none of the properties fall under the two family conversion building type so this can be removed from consideration. While investigating the data further, it was also observed that the majority of properties ranged from 250 total sq ft. to 4000 sq ft. There were 5 properties with over 4000 total sq ft. which were identified as outliers and removed from the sample. This leaves us with a dataset of 85 variables and 2395 properties.

Section 2: Predictive Modeling Framework

For this assignment, we will use a uniform random number to the split the data into a 70/30 train/test split. This split will result in two data sets: one for in-sample model development and

the other for testing on out-of-sample data. Here is a breakdown of number of observations in the train/test split:

```
> dim(train.df)
[1] 1686   27
> dim(test.df)
[1] 708  27
```

There are 1686 observations in the training set and 708 in the test set for a total of 2,394 observations. We will 'train' each model by estimating the models on the 70% of the data identified as the training data set, and we will 'test' each model by examining the predictive accuracy on the 30% of the data.

Section 3: Model Identification by Automated Variable Selection

Model specification starts with choosing the predictor variables. For this assignment, we will use exploratory data analysis, results from previous assignments as well as common sense about housing value to select predictor variables of interest. In addition to the variables created in previous assignments, two variables “QualityIndex” and “TotalSqftCalc” will be used for future modeling purposes. EDA is the basic building block for visually observing the data so we can make a good judgment and compare to the statistics generated at each step in model building process. Here is a list of the predictor variables of interest:

```
"SalePrice"              "LotArea"          "OverallQual"
"OverallCond"            "GarageArea"       "KitchenQual"
"Neighborhood_subgroups" "HouseAge"         "BsmtQual"
"GarageCars"             "PavedDrive"       "TotalFloorSF"
"SaleCondition"          "price_sqft"       "GarageFinish"
"FullBath"               "BsmtFinSF1"       "BsmtFinSF2"
"GrLivArea"              "QualityIndex"     "TotalSqftCalc"
```

Forward Model:

In the forward variable selection method, the model starts with intercept model and one variable is added at a time until optimal model is determined based on fit statistics. The 70% data is fit using forward selection algorithm where 13 predictor variables are analyzed one at a time (R

automatically converts categorical variables into dummy variables). The final model with minimum AIC value for predicting the SalePrice is determined as (forward.lm):
The overall F-statistic indicates the model is significant. The model achieved Adjusted R-squared value of 97.62% which is close to the R-sq value of 97.66%. This indicates that the variables selected are capturing almost all possible variation in the model. The residual standard error indicates a possible predictive error in home sale prices is +/- $12420. Most of the predictors are statistically significant based on their t-values, except some predictor variables where it may be possible to re-group them.

```
Call:
lm(formula = SalePrice ~ OverallQual + TotalSqftCalc + price_sqft +
    TotalFloorSF + BsmtQual + Neighborhood_subgroups + KitchenQual +
    OverallCond + PavedDrive + LotArea + SaleCondition + FullBath +
    QualityIndex, data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-87880  -4930    318   5095  89169

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.646e+05  8.598e+03 -19.148  < 2e-16 ***
OverallQual                -1.257e+03  1.302e+03  -0.966 0.334410
TotalSqftCalc               2.803e+00  8.374e-01   3.347 0.000836 ***
price_sqft                  1.787e+03  3.013e+01  59.302  < 2e-16 ***
TotalFloorSF                1.179e+02  1.688e+00  69.873  < 2e-16 ***
BsmtQualFa                 -1.633e+03  2.663e+03  -0.613 0.539888
BsmtQualGd                 -1.281e+04  1.446e+03  -8.861  < 2e-16 ***
BsmtQualTA                 -1.125e+04  1.739e+03  -6.469 1.31e-10 ***
Neighborhood_subgroupsgrp2 -2.099e+03  1.184e+03  -1.773 0.076427 .
Neighborhood_subgroupsgrp3 -8.212e+03  1.522e+03  -5.396 7.85e-08 ***
Neighborhood_subgroupsgrp4 -1.909e+04  2.175e+03  -8.775  < 2e-16 ***
KitchenQualFa              -8.840e+03  2.748e+03  -3.217 0.001323 **
KitchenQualGd              -1.110e+04  1.510e+03  -7.352 3.11e-13 ***
KitchenQualPo              -3.973e+03  1.268e+04  -0.313 0.754081
KitchenQualTA              -1.067e+04  1.690e+03  -6.316 3.48e-10 ***
OverallCond                -5.192e+03  1.323e+03  -3.926 9.01e-05 ***
PavedDriveP                -4.014e+02  2.417e+03  -0.166 0.868150
PavedDriveY                -4.712e+03  1.518e+03  -3.105 0.001938 **
LotArea                     1.562e-01  4.796e-02   3.256 0.001152 **
SaleConditionAdjLand        2.714e+03  5.797e+03   0.468 0.639695
SaleConditionAlloca        -5.554e+03  4.956e+03  -1.121 0.262541
SaleConditionFamily        -3.362e+03  2.823e+03  -1.191 0.233768
SaleConditionNormal        -1.352e+03  1.380e+03  -0.980 0.327364
SaleConditionPartial        3.229e+03  1.847e+03   1.748 0.080689 .
FullBath                   -2.247e+03  8.721e+02  -2.576 0.010075 *
QualityIndex                5.351e+02  2.201e+02   2.431 0.015157 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12420 on 1582 degrees of freedom
Multiple R-squared:  0.9766,    Adjusted R-squared:  0.9762
F-statistic:  2636 on 25 and 1582 DF,  p-value: < 2.2e-16
```

Backward Model:

In the backward variable elimination method, the model starts with all potential predictor variables and one variable is dropped at a time until an optimal model is determined based on fit statistics. This model includes 12 predictor variables. 70% of the data is fit using the backward elimination algorithm (once again, R automatically converts categorical variables into dummy variables). The final model with minimum AIC value for predicting the SalePrice is as follows:

```
Call:
lm(formula = SalePrice ~ LotArea + OverallCond + KitchenQual +
    Neighborhood_subgroups + BsmtQual + PavedDrive + TotalFloorSF +
    SaleCondition + price_sqft + FullBath + QualityIndex + TotalSqftCalc,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-89849  -4938    351   5133  88911

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.714e+05  4.983e+03 -34.398  < 2e-16 ***
LotArea                     1.603e-01  4.777e-02   3.356 0.000811 ***
OverallCond                -4.022e+03  5.293e+02  -7.600 5.06e-14 ***
KitchenQualFa              -8.732e+03  2.746e+03  -3.180 0.001501 **
KitchenQualGd              -1.108e+04  1.510e+03  -7.343 3.33e-13 ***
KitchenQualPo              -3.502e+03  1.267e+04  -0.276 0.782315
KitchenQualTA              -1.059e+04  1.688e+03  -6.277 4.46e-10 ***
Neighborhood_subgroupsgrp2 -2.074e+03  1.183e+03  -1.753 0.079838 .
Neighborhood_subgroupsgrp3 -8.177e+03  1.521e+03  -5.375 8.81e-08 ***
Neighborhood_subgroupsgrp4 -1.901e+04  2.174e+03  -8.745  < 2e-16 ***
BsmtQualFa                 -1.353e+03  2.647e+03  -0.511 0.609353
BsmtQualGd                 -1.262e+04  1.432e+03  -8.812  < 2e-16 ***
BsmtQualTA                 -1.093e+04  1.707e+03  -6.403 2.01e-10 ***
PavedDriveP                -4.589e+02  2.417e+03  -0.190 0.849425
PavedDriveY                -4.815e+03  1.514e+03  -3.181 0.001498 **
TotalFloorSF                1.177e+02  1.666e+00  70.618  < 2e-16 ***
SaleConditionAdjLand        2.627e+03  5.797e+03   0.453 0.650408
SaleConditionAlloca        -5.628e+03  4.955e+03  -1.136 0.256235
SaleConditionFamily        -3.472e+03  2.820e+03  -1.231 0.218468
SaleConditionNormal        -1.395e+03  1.379e+03  -1.011 0.311964
SaleConditionPartial        3.162e+03  1.846e+03   1.713 0.086964 .
price_sqft                  1.781e+03  2.961e+01  60.170  < 2e-16 ***
FullBath                   -2.220e+03  8.716e+02  -2.547 0.010953 *
QualityIndex                3.347e+02  7.344e+01   4.558 5.56e-06 ***
TotalSqftCalc               2.866e+00  8.348e-01   3.433 0.000613 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12420 on 1583 degrees of freedom
Multiple R-squared:  0.9765,    Adjusted R-squared:  0.9762
F-statistic:  2746 on 24 and 1583 DF,  p-value: < 2.2e-16
```

The backwards model achieved an Adjusted R-squared value of 97.62% and an R-squared value of 97.65% meaning almost all of the possible variation in SalePrice is captured by the model. The residual standard error also indicates a possible predictive error in home sale prices as +/- $12420, same as the forward model.

Stepwise Model:

Stepwise variable selection method is essentially same as forward selection method with an additional provision of being able to delete the variable selected in the previous steps. This model chose 12 predictor variables. The final model with minimum AIC value for predicting the Saleprice is determined using stepwise selection method as (stepwise.lm):

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + price_sqft + BsmtQual +
    Neighborhood_subgroups + KitchenQual + OverallCond + QualityIndex +
    LotArea + PavedDrive + TotalSqftCalc + SaleCondition + FullBath,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-89849  -4938    351   5133  88911

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.714e+05  4.983e+03 -34.398  < 2e-16 ***
TotalFloorSF               1.177e+02  1.666e+00  70.618  < 2e-16 ***
price_sqft                 1.781e+03  2.961e+01  60.170  < 2e-16 ***
BsmtQualFa                -1.353e+03  2.647e+03  -0.511 0.609353
BsmtQualGd                -1.262e+04  1.432e+03  -8.812  < 2e-16 ***
BsmtQualTA                -1.093e+04  1.707e+03  -6.403 2.01e-10 ***
Neighborhood_subgroupsgrp2 -2.074e+03  1.183e+03  -1.753 0.079838 .
Neighborhood_subgroupsgrp3 -8.177e+03  1.521e+03  -5.375 8.81e-08 ***
Neighborhood_subgroupsgrp4 -1.901e+04  2.174e+03  -8.745  < 2e-16 ***
KitchenQualFa             -8.732e+03  2.746e+03  -3.180 0.001501 **
KitchenQualGd             -1.108e+04  1.510e+03  -7.343 3.33e-13 ***
KitchenQualPo             -3.502e+03  1.267e+04  -0.276 0.782315
KitchenQualTA             -1.059e+04  1.688e+03  -6.277 4.46e-10 ***
OverallCond               -4.022e+03  5.293e+02  -7.600 5.06e-14 ***
QualityIndex               3.347e+02  7.344e+01   4.558 5.56e-06 ***
LotArea                    1.603e-01  4.777e-02   3.356 0.000811 ***
PavedDriveP               -4.589e+02  2.417e+03  -0.190 0.849425
PavedDriveY               -4.815e+03  1.514e+03  -3.181 0.001498 **
TotalSqftCalc              2.866e+00  8.348e-01   3.433 0.000613 ***
SaleConditionAdjLand       2.627e+03  5.797e+03   0.453 0.650408
SaleConditionAlloca       -5.628e+03  4.955e+03  -1.136 0.256235
SaleConditionFamily       -3.472e+03  2.820e+03  -1.231 0.218468
SaleConditionNormal       -1.395e+03  1.379e+03  -1.011 0.311964
SaleConditionPartial       3.162e+03  1.846e+03   1.713 0.086964 .
FullBath                  -2.220e+03  8.716e+02  -2.547 0.010953 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12420 on 1583 degrees of freedom
Multiple R-squared:  0.9765,    Adjusted R-squared:  0.9762
F-statistic:  2746 on 24 and 1583 DF,  p-value: < 2.2e-16
```

This model achieved an adjusted R-squared value of 97.62% and an R-squared value of 97.65%. The residual standard error also indicates a possible predictive error in home sale prices as +/- $12420, same as the forward model and the backward model

Junk Model:

Another model is fit using two variables: OverallQual and LotArea. This model is the "Junk Model" and is used as a benchmark for comparison with the other models that use automated variable selection methods. The model is fitted using 70% of training data as:

```
Call:
lm(formula = SalePrice ~ OverallQual + LotArea, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max
-301895  -27639   -2659   21449  288933

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.211e+05  5.360e+03  -22.59   <2e-16 ***
OverallQual  4.614e+04  8.387e+02   55.02   <2e-16 ***
LotArea      2.207e+00  1.605e-01   13.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45040 on 1605 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6867
F-statistic:  1762 on 2 and 1605 DF,  p-value: < 2.2e-16
```

The Adjusted R-Squared value is 68.67% and the R-Squared value is 68.71%. The residual standard error also indicates a possible predictive error in home sale prices as +/- $45040. This model is clearly inferior to the other models.

Multicollinearity:

One area of concern in multiple variable regression is correlation/interactions between predictor variables (known as collinearity). If not addressed properly, it can lead to inaccurate models that don't reflect the data. Variation inflation factor (VIF) is a statistical metric calculated using each predictor as a response. If the VIF of any predictor variable is high (>10 in this case), it may indicate that particular variable is correlated with other predictor variables. Below are the VIF scores for each model:

Forward Model VIF:

```
> sort(vif(forward.lm),decreasing=TRUE)
 [1] 41.142279 32.426637 22.365493  8.283511  7.124168  6.540079  6.414225  5.694439  5.000000  4.729217  4.230106  4.000000  3.813964  3.000000
[15]  3.000000  2.887575  2.878109  2.669114  2.325939  2.000000  1.952937  1.539102  1.525103  1.367510  1.296260  1.271721  1.202366  1.141739
[29]  1.096524  1.067021  1.044063  1.000000  1.000000  1.000000  1.000000  1.000000  1.000000  1.000000  1.000000
```

The first three VIF scores indicate that some variables are correlated with other predictor variables. These variables are OverallQual, TotalSqftCalc and price_sqft. Removing these predictor variables from the model improves VIF but has a negative effect on the R-squared value and residual standard error. Below are the updated VIF values as well as comparison of the three models (Forward, Backward, Stepwise):

```
> sort(vif(forward.lm),decreasing=TRUE)
 [1] 6.344009 5.155356 5.000000 4.504856 4.295493 4.000000 3.808641 3.458059 3.246905 3.000000 3.000000 2.868498 2.518732 2.406343 2.122465 2.072557
[17] 2.000000 1.962990 1.951574 1.859586 1.551239 1.501155 1.413342 1.401067 1.314346 1.216872 1.176956 1.140793 1.090340 1.084876 1.041460 1.000000
[33] 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
```

Observations	1,608	1,608	1,608
R2	0.926	0.977	0.977
Adjusted R2	0.925	0.976	0.976
Residual Std. Error	22,016.140 (df = 1582)	12,415.810 (df = 1583)	12,415.810 (df = 1583)
F Statistic	795.359*** (df = 25; 1582)	2,746.407*** (df = 24; 1583)	2,746.407*** (df = 24; 1583)
Note:			*p<0.1; **p<0.05; ***p<0.01

Backward Model VIF:

```
> sort(vif(backward.lm),decreasing=TRUE)
 [1] 7.997812 6.944380 6.529202 5.000000 4.582108 4.064919 4.000000 3.790862 3.581769 3.000000 3.000000 2.864745 2.828040 2.635219 2.323622 2.140586
[17] 2.000000 1.947014 1.892556 1.535445 1.524343 1.367131 1.289553 1.263306 1.192871 1.140606 1.092187 1.065638 1.043815 1.000000 1.000000 1.000000
[33] 1.000000 1.000000 1.000000 1.000000
```

There does not appear to be any collinearity between variables that needs to be removed.

Stepwise Model VIF:

```
> sort(vif(stepwise.lm),decreasing=TRUE)
 [1] 7.997812 6.944380 6.529202 5.000000 4.582108 4.064919 4.000000 3.790862 3.581769 3.000000 3.000000 2.864745 2.828040 2.635219 2.323622 2.140586
[17] 2.000000 1.947014 1.892556 1.535445 1.524343 1.367131 1.289553 1.263306 1.192871 1.140606 1.092187 1.065638 1.043815 1.000000 1.000000 1.000000
[33] 1.000000 1.000000 1.000000 1.000000
```

There does not appear to be any collinearity between variables that needs to be removed.

Junk Model VIF:

```
> sort(vif(junk.lm),decreasing=TRUE)
OverallQual     LotArea
   1.023287    1.023287
```

There does not appear to be any collinearity between variables that needs to be removed.

Model Comparison:

Here is a comparison of the models based on 5 metrics (Adj R-Squared, AIC, BIC, MSE and MAE). Based on the results, either the Backward or Stepwise models are most suitable for the data. They are the best performers for all 5 metrics.

Table: Model Fit Metrics Comparison

Model	adj.r.squared	AIC	BIC	MSE	MAE
Junk	0.6867	39028.56	39050.10	2024610747	32185.564
Forward Selection	0.9251	36749.59	36894.92	476873177	14406.895
Backward Elimination	0.9762	34906.46	35046.41	151755688	7971.034
Stepwise	0.9762	34906.46	35046.41	151755688	7971.034

Table: Model Fit Metrics Ranks

Model	Rank.adjR	Rank.AIC	Rank.BIC	Rank.MSE	Rank.MAE
Junk	3	3	3	3	3
Forward Selection	2	2	2	2	2
Backward Elimination	1	1	1	1	1
Stepwise	1	1	1	1	1

Section 4: Predictive Accuracy

As shown below, the four final models are compared against each other in terms of insample model fit and predictive accuracy for out-of-sample data. The metrics used are MAE (Mean Absolute Error) and MSE (Mean Squared Error(MSE). These metrics are more attuned to measuring how good the model is at predicting home-price values. On the other hand, metrics such as adjusted r-sq, AIC, BIC metrics focus on amount of variation captured by the models. Therefore, between the metrics, measures of both accuracy and precision are analyzed. Based on these results, the Backward and Stepwise models performed best.

Table: Prediction Metrics Comparision

Model	MAE.In.Sample	MAE.Out.of.Sample	MSE.In.Sample	MSE.Out.of.Sample
Junk	32185.564	30072.541	2024610747	1858994119
Forward	14406.895	14271.827	476873177	426238970
Backward	7971.034	8063.802	151755688	149364323
Stepwise	7971.034	8063.802	151755688	149364323

Section 5: Operational Validation

We have chosen the Stepwise Model to evaluate further. Although we have validated the model in a statistical sense, we still need to see how this translates to real-world policy. For example, we may want to define a cut-off point in which we want a certain level of accuracy. Below are grades for the Stepwise model in training and in testing. We will 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. According to these scores about 89% of the predictions lie within 10% of the actual values. For business purposes, I would say that this is a good grade although there is still room for improvement.

Training Grade:

```
stepwise.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.88557214           0.05721393           0.04166667           0.01554726
```

Test Grade:

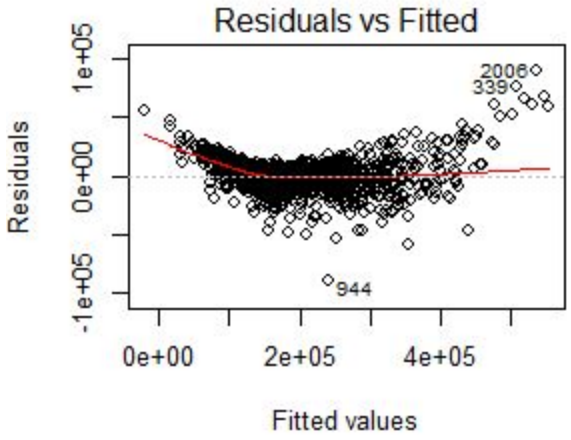
```
stepwise.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.89160305           0.05954198           0.03206107           0.01679389
```

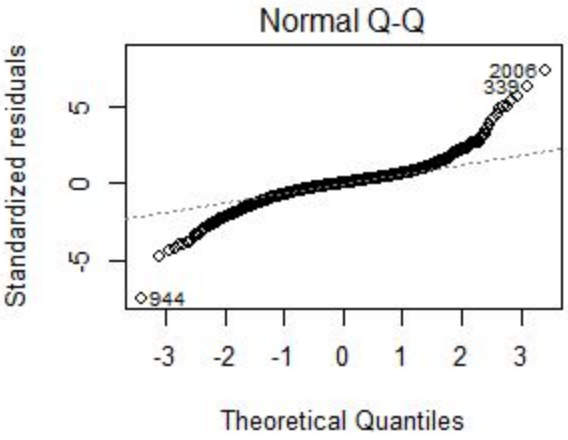
Section 6: Final Model

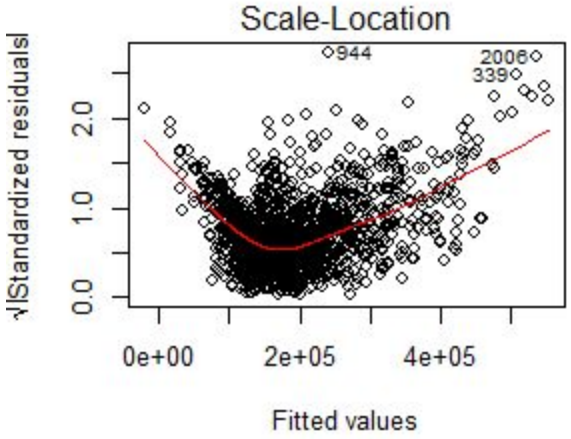
In this section, we will review some diagnostic plots for the model. Looking at the visualizations below, it appears that there are outliers that may be skewing the data and causing the residuals to violate assumptions of normality. We will remove those points and take the log of SalePrice to combat these errors. Below is the the initial model with plots followed by the the final model with plots.

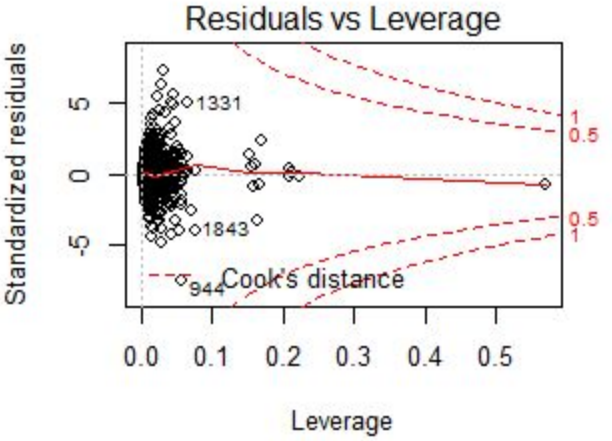
Initial Model:

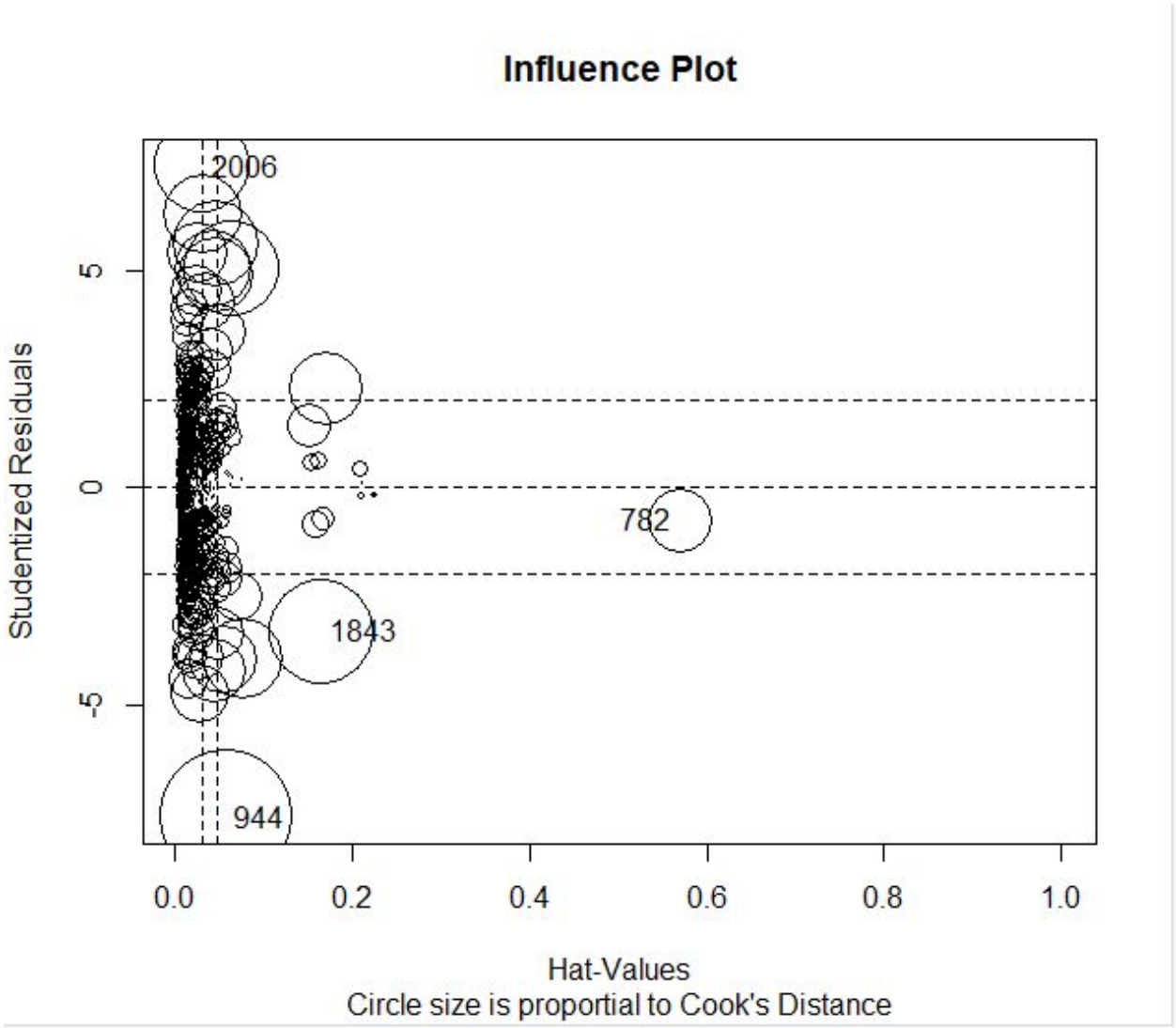
```
Call:
lm(formula = SalePrice ~ TotalFloorSF + price_sqft + BsmtQual +
    Neighborhood_subgroups + KitchenQual + OverallCond + QualityIndex +
    LotArea + PavedDrive + TotalSqftCalc + SaleCondition + FullBath,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-89849  -4938    351   5133  88911

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.714e+05  4.983e+03 -34.398  < 2e-16 ***
TotalFloorSF                1.177e+02  1.666e+00  70.618  < 2e-16 ***
price_sqft                  1.781e+03  2.961e+01  60.170  < 2e-16 ***
BsmtQualFa                 -1.353e+03  2.647e+03  -0.511 0.609353
BsmtQualGd                 -1.262e+04  1.432e+03  -8.812  < 2e-16 ***
BsmtQualTA                 -1.093e+04  1.707e+03  -6.403 2.01e-10 ***
Neighborhood_subgroupsgrp2 -2.074e+03  1.183e+03  -1.753 0.079838 .
Neighborhood_subgroupsgrp3 -8.177e+03  1.521e+03  -5.375 8.81e-08 ***
Neighborhood_subgroupsgrp4 -1.901e+04  2.174e+03  -8.745  < 2e-16 ***
KitchenQualFa              -8.732e+03  2.746e+03  -3.180 0.001501 **
KitchenQualGd              -1.108e+04  1.510e+03  -7.343 3.33e-13 ***
KitchenQualPo              -3.502e+03  1.267e+04  -0.276 0.782315
KitchenQualTA              -1.059e+04  1.688e+03  -6.277 4.46e-10 ***
OverallCond                -4.022e+03  5.293e+02  -7.600 5.06e-14 ***
QualityIndex                3.347e+02  7.344e+01   4.558 5.56e-06 ***
LotArea                     1.603e-01  4.777e-02   3.356 0.000811 ***
PavedDriveP                -4.589e+02  2.417e+03  -0.190 0.849425
PavedDriveY                -4.815e+03  1.514e+03  -3.181 0.001498 **
TotalSqftCalc               2.866e+00  8.348e-01   3.433 0.000613 ***
SaleConditionAdjLand        2.627e+03  5.797e+03   0.453 0.650408
SaleConditionAlloca        -5.628e+03  4.955e+03  -1.136 0.256235
SaleConditionFamily        -3.472e+03  2.820e+03  -1.231 0.218468
SaleConditionNormal        -1.395e+03  1.379e+03  -1.011 0.311964
SaleConditionPartial        3.162e+03  1.846e+03   1.713 0.086964 .
FullBath                   -2.220e+03  8.716e+02  -2.547 0.010953 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12420 on 1583 degrees of freedom
Multiple R-squared:  0.9765,    Adjusted R-squared:  0.9762
F-statistic:  2746 on 24 and 1583 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
2006
339
944
1e+05
0e+00
-1e+05
0e+00
2e+05
4e+05


Normal Q-Q
Standardized residuals
Theoretical Quantiles
2006
339
944
5
0
-5
-3
-2
-1
0
1
2
3

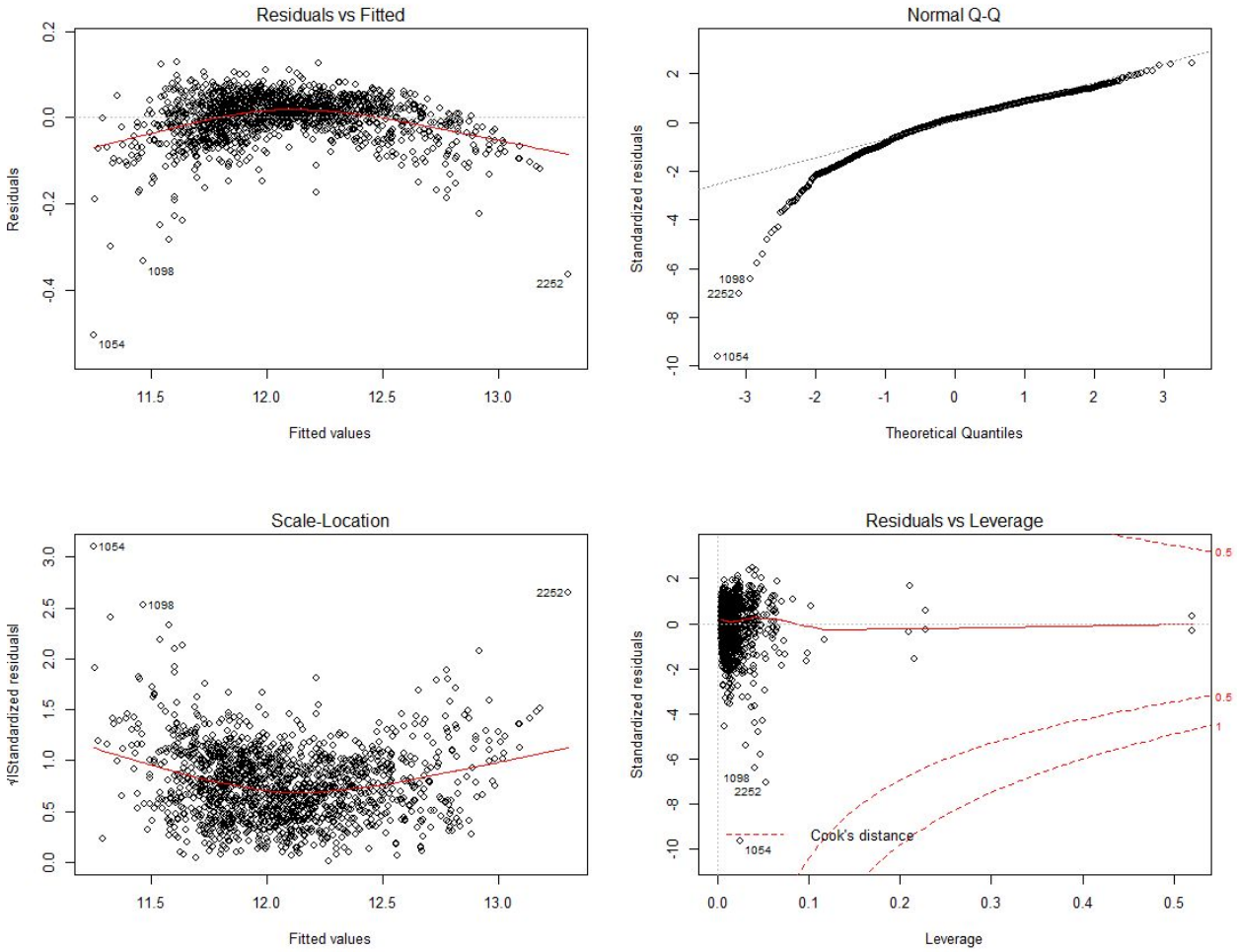

Scale-Location
√|Standardized residuals|
Fitted values
944
2006
339
2.0
1.0
0.0
0e+00
2e+05
4e+05


Residuals vs Leverage
Standardized residuals
Leverage
1331
1843
944
Cook's distance
1
0.5
0.5
1
5
0
-5
0.0
0.1
0.2
0.3
0.4
0.5


Influence Plot
2006
782
1843
944
Studentized Residuals
5
0
-5
0.0
0.2
0.4
0.6
0.8
1.0
Hat-Values
Circle size is proportial to Cook's Distance

Final Model:

```
Call:
lm(formula = log(SalePrice) ~ TotalFloorSF + price_sqft + BsmtQual +
    Neighborhood_subgroups + KitchenQual + OverallCond + QualityIndex +
    LotArea + PavedDrive + TotalSqftCalc + SaleCondition + FullBath,
    data = subdatinf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50385 -0.02244  0.00885  0.03283  0.12721

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.016e+01  2.518e-02 403.350  < 2e-16 ***
TotalFloorSF                5.570e-04  7.932e-06  70.224  < 2e-16 ***
price_sqft                  7.022e-03  1.524e-04  46.068  < 2e-16 ***
BsmtQualFa                 -6.692e-02  1.260e-02  -5.311 1.25e-07 ***
BsmtQualGd                  1.460e-02  6.648e-03   2.196 0.028243 *
BsmtQualTA                 -8.852e-03  7.884e-03  -1.123 0.261723
Neighborhood_subgroupsgrp2  3.193e-02  5.417e-03   5.894 4.66e-09 ***
Neighborhood_subgroupsgrp3  2.492e-02  7.203e-03   3.460 0.000556 ***
Neighborhood_subgroupsgrp4 -3.847e-03  1.050e-02  -0.367 0.714041
KitchenQualFa              -1.133e-02  1.278e-02  -0.887 0.375482
KitchenQualGd               2.484e-02  7.144e-03   3.477 0.000522 ***
KitchenQualTA               1.273e-02  7.917e-03   1.607 0.108178
OverallCond                -1.625e-02  2.409e-03  -6.746 2.17e-11 ***
QualityIndex                3.206e-03  3.340e-04   9.598  < 2e-16 ***
LotArea                     1.477e-06  3.432e-07   4.305 1.78e-05 ***
PavedDriveP                 2.906e-02  1.117e-02   2.602 0.009366 **
PavedDriveY                 4.094e-02  7.233e-03   5.661 1.80e-08 ***
TotalSqftCalc               1.703e-05  3.736e-06   4.557 5.60e-06 ***
SaleConditionAdjLand        2.909e-02  2.506e-02   1.161 0.245900
SaleConditionAlloca         8.479e-02  3.877e-02   2.187 0.028889 *
SaleConditionFamily         2.640e-02  1.394e-02   1.894 0.058457 .
SaleConditionNormal         3.284e-02  6.308e-03   5.207 2.19e-07 ***
SaleConditionPartial        4.285e-02  8.399e-03   5.102 3.80e-07 ***
FullBath                    1.933e-02  3.920e-03   4.931 9.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05305 on 1488 degrees of freedom
Multiple R-squared:  0.9769,    Adjusted R-squared:  0.9765
F-statistic:  2731 on 23 and 1488 DF,  p-value: < 2.2e-16
```



Section 7: Reflection/Conclusion

- Defining the sample population is very important because one may be answering the wrong question if this is not well defined
- EDA is both an art and a science. There are certain steps/guidelines to follow, but there are tricks the data scientist can do to really get to know the data. This requires lots of time spent with the data.
- Automated variables selection is a useful trick but it is not a completely automated process. Model validation is still an important aspect of this process
- It is important to keep in mind what the overall business question is. It is easy to just get caught up in the numbers and follow them blindly. Every calculation in this process needs to have the overall goal in mind.

Tips to improve predictive accuracy:

- The final model ended up having predictions within 10% of the actual observations 89% of the time in training as well as in the test set. I would say that this is a good start but there is still more to do in terms of dealing with outliers and OLS assumptions.
- There are so many ways to build a model and exhausting all the possibilities would take lots of time and effort. Continuing the process of automatic variable selection on another pool of variables could improve the final model.