Introduction

The purpose of this document is to develop an ideal regression model to predict home sale price using variables in the AMES data. The response variable is SalePrice (Y). Specific models will be developed and diagnostics will be reviewed to understand the fit.

Part 1

Initial review of the AMES data revealed a number of null values as well as empty values. In some cases the null or missing values were dropped or replaced with a median or mean value. From the original data, 2927 observations remain, with 87 variables including SalePrice; the original dataset included 2930 observations; only 3 were dropped. Five new variables have been created and are shared in Appendix 1.

Prior to building a model, it is important to consider the categorical variables that may serve as quality explanatory variables. Based on initial review of the data, and using general logic, the most relevant categorical variables that may be valuable in explaining SalePrice are: Zoning, Utilities, and BldgType. OverallQual and OverallCond are also categorical, but they're represented as ordinal variables, so they will be excluded for comparison by factor level.

Boxplots, means and standard deviations are captured for Zoning, Utilities, and BldgType in Appendix 2. The difference in means is important to consider as the factor levels will help to explain the dependent variable SalePrice. Utilities is the most distinct variable, with a significant difference in the mean value of SalePrice by factor. Zoning and BldgType also contain differentiation between factors, however there is some overlap in factors. Utilities only contains one instance of the "NoSeWa" factor level, so this instance has been dropped from the dataset. The Utilities variable can now be converted to a two factor classification variable (0,1). However, after further review, only 2 instances of the 0 category exist for Utilities. This is not a sufficient number of instances, so Utilities should not be further considered. The standard deviations by factor also show the variance within each factor level. Larger standard deviations eludes to less differentiation between factors – there will be some overlap. Zoning and BldgType are the two remaining categorical variables of primary interest.

Appendix 2 provides summaries of regression models for the Zoning and BldgType dummy variables; in each case one variable is left out. The Zoning model has an r^2 of ~.11, while the BldgType model has an r^2 of ~.03. Both are statistically significant models. The conclusion is that the Zoning variables explain SalePrice more effectively.

Part 2

The dataset has been split into a training set and a test set, 70/30. A table of the counts is provided below:

Dataframe	Observations
train.df	2047
test.df	879
mydata	2926

The split will allow us to fit models to the training data and validate the model on the test data. This will help further confirm a model's reliability.

Part 3

Moving forward the following variables will be considered for a final model: QualityIndex, TotalSqftCalc, OQ_SF, Zone_C, Zone_FV, Zone_I, Zone_RH, Zone_RL, Zone_RM, OverallQual, Overal_Cond, OQ_Rooms, OC_SF, GrLivArea, and YearBuilt.

Using the stepAIC() function, the results of the forward, backward, and stepwise models are provided in Appendix 3. An initial observation is that multicollinearity will exist within each of the models, so further scrutiny on the explanatory variables will be required. Examples of pairs with correlation include OC_SF and TotalSqftCalc, OverallQual and QualityIndex, and TotalSqftCalc and GrLivArea.

Based on the VIF values, OQ_SF and TotalSqftCalc should be removed from the explanatory variables. The VIF values are presented in Appendix 3. The resulting updated models are also provided. Below is a table of the coefficients and corresponding VIF values for each model.

Forward	VIF	Backward	VIF	Stepwise	VIF	Junk	VIF
(Intercept) -885431.7901		(Intercept) -888841.4853		(Intercept) -885431.79		(Intercept) -172338.624	
OQ_Rooms 823.5195	7.122304	OverallQual 28079.0046	7.849119	QualityIndex -1643.95	4.557325	OverallQual 40777.179	23.213317
OverallQual 27878.7571	7.761173	YearBuilt 427.8617	2.450072	OQ_Rooms 823.5195	7.122304	OverallCond 11871.057	17.395051
OC_SF 6.9209	3.321212	GrLivArea 13.9455	5.691086	YearBuilt 417.9492	2.416485	QualityIndex -2255.106	36.483731
QualityIndex -1643.9529	4.557325	OC_SF 6.8936	3.392732	OC_SF 6.9209	3.321212	GrLivArea 20.942	2.889410
YearBuilt 417.9492	2.416485	OQ_Rooms 803.8430	7.284674	OverallQual 27878.7571	7.761173	TotalSqftCalc 41.911	2.784967
GrLivArea 13.6710	5.604607	Zone_FV -21234.9522	5.657088	GrLivArea 13.6710	5.604607		

Zone_RM -7650.6346	1.299086	Zone_RH -38705.859	1.898232	Zone_RM -7650.6346	1.299086		
Zone_RH -21633.1393	1.011023	Zone_RL -17259.4542	16.687389	Zone_RH -21633.1393	1.011023		
Zone_C 12143.9570	1.069165	Zone_RM -24656.8751	12.717017	Zone_C 12143.9570	1.069165		
		QualityIndex -1624.519	4.571808				

Comparison of fit and metrics, model ranks by category in parenthesis:

Model	Forward	Backward	Stepwise	Junk
r^2	0.8212 (2)	0.8215 (1)	0.8212 (2)	0.8036 (3)
AIC	42770.32 (2)	42768.05 (1)	42770.32 (2)	42954.06 (3)
BIC	48643.32 (1)	48646.67 (2)	48643.32 (1)	48804.56 (3)
MSE	1180581343 (2)	1178700520 (1)	1180581343 (2)	1293970882 (3)
MAE	23893.75 (1)	23897.52 (2)	23893.75 (1)	24978.58 (3)

As seen in the table above, the forward and stepwise models are identical; however, the backward model outperforms all other models in 3 of 5 categories.

Part 4

Model	Forward	Backward	Stepwise	Junk
MSE	1695184104	1696576678	1695184104	1881283694
MAE	23659.15	23703.31	23659.15	24750.36

Based on the metrics provided in the table above, the forward and stepwise models fit the best. The backward model outperformed in fitting, considering the 5 metrics evaluated. Based on this experiment, BIC and MAE are valuable in grasping which models may perform best. When a model has better accuracy out of sample, it may have been underfit. If it has better accuracy in sample, then it may have been overfit. In this case, the forward/stepwise models may have been underfit.

Part 5

Prediction grades are provided in Appendix 4. Based on the prediction grades, the backward model now outperforms the other models, though it is by a small margin. So we have again shifted to the backward model being the best model; though the forward and stepwise models performed better when considering the predictive accuracy results.

When considering the GSEs rating of a model as underwriting quality, all of the models are considered underwriting quality as they are all accurate to within 10% more than 50% of the time.

Part 6

The backward model is selected as the final model. While between the top performing models all are nearly the same in terms of explanatory variables, the backward model has an additional categorical variable. The OverallQual variable should be dropped from the model. It's coefficient is negative in all models and this is not logical.

An initial fit was conducted for the remaining backward model as such: lm(SalePrice~OverallQual+YearBuilt+GrLivArea+OC_SF+OQ_Rooms+Zone_C+Zone_RM+Zone_RL+Zone_I +Zone_FV+Zone_RH, data = train.clean)

When reviewing the coefficients, and as all Zoning dummy variables must remain, an anomaly stood out. The Zone_I variable (industrial zone) coefficient is positive, this does not make sense when considering that the commercial zone variable is negative as is the case with all of the related dummy variables. The r^2 for the model with the dummy variables is 0.8141, the r^2 without them is 0.8108. Therefore, the Zoning dummy variables are removed from the model. It is also worth noting that each other dummy variable had negative coefficients.

The final model is as such: lm(SalePrice~OverallQual+YearBuilt+GrLivArea+OC_SF+OQ_Rooms, data = train.clean). Summary statistics of the model as well as supporting charts and other information are found in Appendix 5. The equation for the model is as follows: SalePrice = 17871.1406(OverallQual) + 625.8537(YearBuilt) + 27.5187(GrLivArea) + 5.0477(OC_SF) + 790.0651(OQ_Rooms) – 1289924.3504. The interpretation of the equation is as follows – for each increase in the level of OverallQual, SalePrice increases by 17871.1406, for each YearBuilt, or as the property becomes newer, SalePrice increases by 625.8537, for each increase in GrLivArea, or for each additional sqft of living area, SalePrice increases by 27.5187, for each unit of the overall condition and total sqft interaction, SalePrice increases by 5.0477, for each unit of the overall quality and total rooms above ground interaction, SalePrice increases by 790.0651; the y-intercept is -1289924.3504. While a property will never be valued at this, the y-intercept is a holding point for the other variables to affect the SalePrice, it is also where all variables equal 0.

The model and all of its variables are statistically significant as found in the model summary and annova results. F-statistic: 1749 on 5 and 2041 DF, p-value: < 0.00000000000000022.

As previously stated, the the r^2 is 0.8108. The MSE is 1246669651. A plot of the residuals is found in Appendix 5. A correlation plot is also provided. The residual plot shows no signs of heteroscedasticity or non-linear patterns. While the correlation plot does raise some concerns, the VIF scores verify that multicollinearity is not an issue; the residual plot also helps to verify the model's effectiveness. The MSE is 1664598806 and the MAE is 24219.33. Prediction grades are also provided in Appendix 5. When considering the GSEs rating of a model as underwriting quality, the final model is considered as underwriting quality as it is accurate to within 10% more than 50% of the time.

Part 7 – Reflection/Conclusions

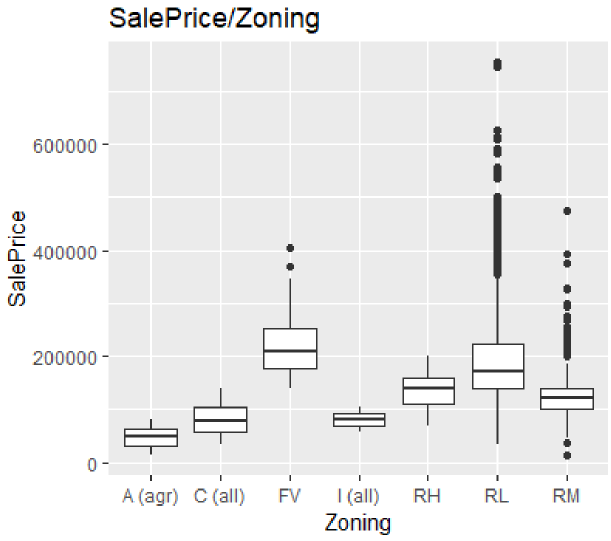
After working with the AMES data, the data does present challenges. First, there are a number of outliers. Another issue is that there are a large amount of variables, many of them being categorical variables which often do not appear relevant. Some of the categorical variables can be subjective, such as OverallQual and OverallCond. These types of variables leave room for error.

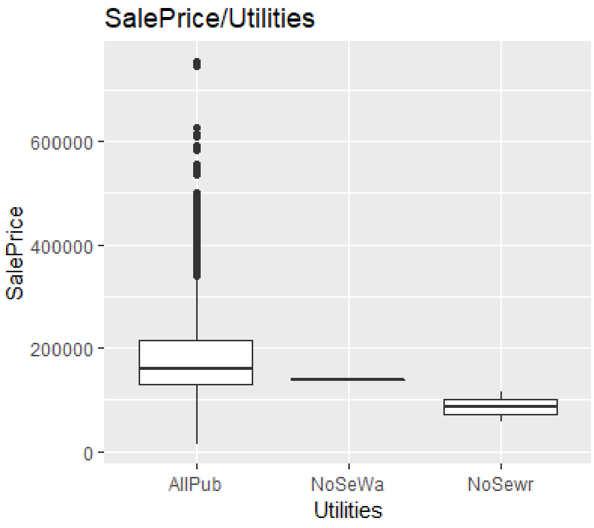
To improve predictive accuracy, it would be worth experimenting with more interactions. In terms of parsimony, I am a firm believer that simpler models are preferred to complicated models. I believe the final model presented here demonstrates that. A simpler, more interpretable model is a better solution as not only do we avoid overfitting, but we also allow room for updates, and it is easier to understand when an outlier may occur; what factors may be driving it.

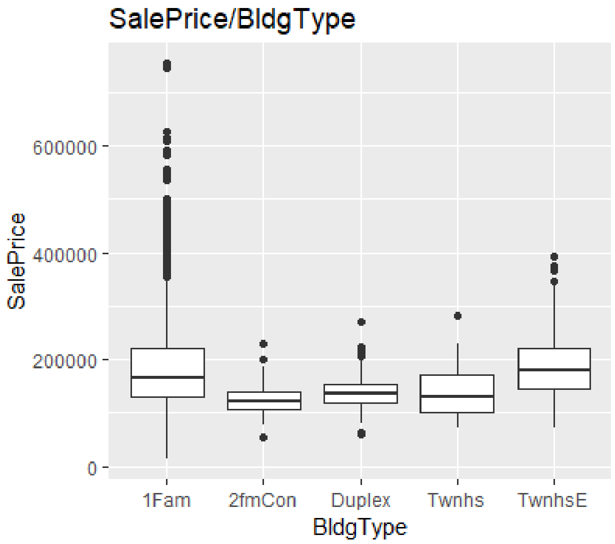
Appendix 1: New Variables

```
mydata$OQ_SF <- mydata$OverallQual * mydata$TotalSqftCalc

mydata$OC_SF <- mydata$OverallCond * mydata$ TotalSqftCalc

mydata$OQ_Rooms <- mydata$OverallQual*mydata$TotRmsAbvGrd

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

mydata$TotalSqftCalc <- mydata$BsmtFinSF1+mydata$BsmtFinSF2+mydata$GrLivArea
```

Appendix 2: Initial Categorical Variable Comparison







Comparison of Means:

```
> mean(mydata$SalePrice[mydata$Zoning == "A (agr)"])
[1] 47300
> mean(mydata$SalePrice[mydata$Zoning == "C (all)"])
[1] 79795.04
> mean(mydata$SalePrice[mydata$Zoning == "FV"])
[1] 218986.9
> mean(mydata$SalePrice[mydata$Zoning == "I (all)"])
[1] 80312.5
> mean(mydata$SalePrice[mydata$Zoning == "RH"])
[1] 136419.8
> mean(mydata$SalePrice[mydata$Zoning == "RL"])
[1] 191293.7
> mean(mydata$SalePrice[mydata$Zoning == "RM"])
[1] 126832.8
```

```
> mean(mydata$SalePrice[mydata$Utilities == "NoSeWa"])
[1] 137500
> mean(mydata$SalePrice[mydata$Utilities == "NoSewr"])
[1] 86312.5
> mean(mydata$SalePrice[mydata$Utilities == "AllPub"])
[1] 180925.1
```

```
> mean(mydata$SalePrice[mydata$BldgType == "1Fam"])
[1] 184876.9
> mean(mydata$SalePrice[mydata$BldgType == "2fmCon"])
[1] 125581.7
> mean(mydata$SalePrice[mydata$BldgType == "Duplex"])
[1] 139808.9
> mean(mydata$SalePrice[mydata$BldgType == "Twnhs"])
[1] 135934.1
> mean(mydata$SalePrice[mydata$BldgType == "TwnhsE"])
[1] 192311.9
```

Comparison of Standard Deviations:

```
> sd(mydata$SalePrice[mydata$Zoning == "A (agr)"])
[1] 48366.1
> sd(mydata$SalePrice[mydata$Zoning == "C (all)"])
[1] 31084.82
> sd(mydata$SalePrice[mydata$Zoning == "FV"])
[1] 52684.23
> sd(mydata$SalePrice[mydata$Zoning == "I (all)"])
[1] 32084.97
> sd(mydata$SalePrice[mydata$Zoning == "RH"])
[1] 36173.1
> sd(mydata$SalePrice[mydata$Zoning == "RL"])
[1] 81312.1
> sd(mydata$SalePrice[mydata$Zoning == "RM"])
[1] 48185.56
```

```
> sd(mydata$SalePrice[mydata$Utilities == "AllPub"])
[1] 79898.33
> sd(mydata$SalePrice[mydata$Utilities == "NoSeWa"])
[1] NA
> sd(mydata$SalePrice[mydata$Utilities == "NoSewr"])
[1] 40570.25
```

```
> sd(mydata$SalePrice[mydata$BldgType == "1Fam"])
[1] 82841.56
> sd(mydata$SalePrice[mydata$BldgType == "2fmCon"])
[1] 31089.24
> sd(mydata$SalePrice[mydata$BldgType == "Duplex"])
[1] 39498.97
> sd(mydata$SalePrice[mydata$BldgType == "Twnhs"])
[1] 41938.93
> sd(mydata$SalePrice[mydata$BldgType == "TwnhsE"])
[1] 66191.74
```

```
> summary(bldg_fit)

Call:
lm(formula = SalePrice ~ Bldg_2Fam + Bldg_Dup + Bldg_Twn + Bldg_TwnE, 
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-172107  -50517  -16434   32104  570104 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   184896       1597 115.789 < 0.0000000000000002 ***
Bldg_2Fam     -59315      10105  -5.870        0.00000000486 ***
Bldg_Dup      -45088       7693  -5.861        0.00000000513 ***
Bldg_Twn      -48962       7980  -6.136        0.00000000096 ***
Bldg_TwnE       7416       5389   1.376                0.169    

Residual standard error: 78570 on 2921 degrees of freedom
Multiple R-squared:  0.03464,	Adjusted R-squared:  0.03332 
F-statistic:  26.2 on 4 and 2921 DF,  p-value: < 0.00000000000000022
```

```
> summary(zone_fit)

Call:
lm(formula = SalePrice ~ Zone_C + Zone_FV + Zone_I + Zone_RH + 
    Zone_RL + Zone_RM, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-156317  -48353  -15035   26183  563683 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    47300      53198   0.889  0.37400    
Zone_C         32495      55285   0.588  0.55673    
```

```
Zone_FV       171687      53579   3.204  0.00137
Zone_I         33013      75233   0.439  0.66084
Zone_RH        89120      55133   1.616  0.10610
Zone_RL       144017      53221   2.706  0.00685
Zone_RM        79533      53313   1.492  0.13586

Residual standard error: 75230 on 2919 degrees of freedom
Multiple R-squared:  0.1155,	Adjusted R-squared:  0.1137
F-statistic: 63.54 on 6 and 2919 DF,  p-value: < 0.00000000000000022
```

Appendix 3: Automatic Variable Selection

```
> summary(forward.lm)

Call:
lm(formula = SalePrice ~ OQ_SF + YearBuilt + OQ_Rooms + TotalSqftCalc + 
    OC_SF + GrLivArea + Zone_RM + OverallQual + Zone_RH + Zone_C, 
    data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-554482  -15027   -1198   13682  184041 

Coefficients:
                  Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)   -887489.9871   62484.4589 -14.203 < 0.0000000000000002 ***
OQ_SF              12.4051       0.5109  24.280 < 0.0000000000000002 ***
YearBuilt         497.9644      32.7026  15.227 < 0.0000000000000002 ***
OQ_Rooms          286.7747     112.0106   2.560               0.0105 *  
TotalSqftCalc     -65.7802       4.5730 -14.384 < 0.0000000000000002 ***
OC_SF               2.9393       0.3346   8.785 < 0.0000000000000002 ***
GrLivArea          26.6364       3.1314   8.506 < 0.0000000000000002 ***
Zone_RM        -12902.9554    2079.8421  -6.204        0.000000000666 ***
OverallQual     -3488.7415    1330.5155  -2.622               0.0088 ** 
Zone_RH        -15385.7475    6630.3722  -2.320               0.0204 *  
Zone_C         -12017.3855    7052.5760  -1.704               0.0885 .  

Residual standard error: 30050 on 2036 degrees of freedom
Multiple R-squared:  0.8632,	Adjusted R-squared:  0.8626 
F-statistic:  1285 on 10 and 2036 DF,  p-value: < 0.00000000000000022
```

```
> summary(backward.lm)

Call:
lm(formula = SalePrice ~ OverallQual + YearBuilt + GrLivArea + 
    OQ_SF + OC_SF + OQ_Rooms + Zone_FV + Zone_RL + TotalSqftCalc, 
    data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-555299  -15052   -1275   13697  183986 

Coefficients:
                  Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)   -899872.9055   62109.6693 -14.488 < 0.0000000000000002
```

```
OverallQual     -3601.7765    1335.7439  -2.696               0.00707
YearBuilt         498.1145      32.8709  15.154 < 0.0000000000000002
GrLivArea          26.8980       3.1553   8.525 < 0.0000000000000002
OQ_SF              12.4661       0.5071  24.581 < 0.0000000000000002
OC_SF               2.9129       0.3339   8.724 < 0.0000000000000002
OQ_Rooms          279.6884     113.1608   2.472               0.01353
Zone_FV         12236.9669    3780.5542   3.237               0.00123
Zone_RL         13060.7559    2000.7108   6.528       0.0000000000839
TotalSqftCalc     -66.1831       4.5570 -14.523 < 0.0000000000000002

Residual standard error: 30050 on 2037 degrees of freedom
Multiple R-squared:  0.8632,	Adjusted R-squared:  0.8626 
F-statistic:  1429 on 9 and 2037 DF,  p-value: < 0.00000000000000022


> summary(stepwise.lm)

Call:
lm(formula = SalePrice ~ TotalSqftCalc + OQ_SF + YearBuilt + 
    GrLivArea + OC_SF + Zone_RM + Zone_RH + OQ_Rooms + OverallQual + 
    Zone_C, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-554482  -15027   -1198   13682  184041 

Coefficients:
                  Estimate   Std. Error t value             Pr(>|t|)
(Intercept)   -887489.9871   62484.4589 -14.203 < 0.0000000000000002
TotalSqftCalc     -65.7802       4.5730 -14.384 < 0.0000000000000002
OQ_SF              12.4051       0.5109  24.280 < 0.0000000000000002
YearBuilt         497.9644      32.7026  15.227 < 0.0000000000000002
GrLivArea          26.6364       3.1314   8.506 < 0.0000000000000002
OC_SF               2.9393       0.3346   8.785 < 0.0000000000000002
Zone_RM        -12902.9554    2079.8421  -6.204        0.000000000666
Zone_RH        -15385.7475    6630.3722  -2.320               0.0204
OQ_Rooms          286.7747     112.0106   2.560               0.0105
OverallQual     -3488.7415    1330.5155  -2.622               0.0088
Zone_C         -12017.3855    7052.5760  -1.704               0.0885

Residual standard error: 30050 on 2036 degrees of freedom
Multiple R-squared:  0.8632,	Adjusted R-squared:  0.8626 
F-statistic:  1285 on 10 and 2036 DF,  p-value: < 0.00000000000000022


> junk.lm <- lm(SalePrice ~ OverallQual + OverallCond + QualityIndex + GrLivA
rea + TotalSqftCalc, data=train.df)
> summary(junk.lm)

Call:
lm(formula = SalePrice ~ OverallQual + OverallCond + QualityIndex + 
    GrLivArea + TotalSqftCalc, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-385907  -20247   -1046   17267  251389 
```

```
Coefficients:
                  Estimate  Std. Error t value            Pr(>|t|)
(Intercept)    -172338.624   15554.149 -11.080 < 0.0000000000000002
OverallQual      40777.179    2695.819  15.126 < 0.0000000000000002
OverallCond      11871.057    2925.970   4.057  0.0000515480058922
QualityIndex     -2255.106     514.921  -4.380  0.0000124963759666
GrLivArea           20.942       2.673   7.835  0.0000000000000751
TotalSqftCalc       41.911       1.783  23.509 < 0.0000000000000002

Residual standard error: 35970 on 2041 degrees of freedom
Multiple R-squared:  0.8036,	Adjusted R-squared:  0.8031
F-statistic:  1670 on 5 and 2041 DF,  p-value: < 0.00000000000000022


> sort(vif(forward.lm),decreasing=TRUE)
        OQ_SF TotalSqftCalc   OverallQual      OQ_Rooms     GrLivArea
OC_SF     YearBuilt       Zone_RM        Zone_C
    30.628726     26.251474      8.100818      7.369316      5.680726      4.
833276      2.253241      1.309388      1.090612
      Zone_RH
     1.011639
> sort(vif(backward.lm),decreasing=TRUE)
        OQ_SF TotalSqftCalc   OverallQual      OQ_Rooms     GrLivArea
OC_SF     YearBuilt       Zone_RL       Zone_FV
    30.192864     26.080748      8.168560      7.525080      5.770634      4.
815511      2.277599      1.610246      1.534376
> sort(vif(stepwise.lm),decreasing=TRUE)
        OQ_SF TotalSqftCalc   OverallQual      OQ_Rooms     GrLivArea
OC_SF     YearBuilt       Zone_RM        Zone_C
    30.628726     26.251474      8.100818      7.369316      5.680726      4.
833276      2.253241      1.309388      1.090612
      Zone_RH
     1.011639
```

After removing OQ_SF and TotalSFCalc:

```
> summary(forward.lm)

Call:
lm(formula = SalePrice ~ OQ_Rooms + OverallQual + OC_SF + QualityIndex + 
    YearBuilt + GrLivArea + Zone_RM + Zone_RH + Zone_C, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-359761  -19513   -3029   16086  245893 

Coefficients:
                 Estimate   Std. Error t value             Pr(>|t|)
(Intercept)  -885431.7901   74287.3402 -11.919 < 0.0000000000000002
OQ_Rooms         823.5195     125.8949   6.541      0.0000000000769
OverallQual    27878.7571    1488.9207  18.724 < 0.0000000000000002
OC_SF              6.9209       0.3171  21.827 < 0.0000000000000002
QualityIndex   -1643.9529     173.8329  -9.457 < 0.0000000000000002
YearBuilt        417.9492      38.7188  10.794 < 0.0000000000000002
GrLivArea         13.6710       3.5560   3.844             0.000125
Zone_RM        -7650.6346    2368.4681  -3.230             0.001257
```

```
Zone_RH       -21633.1393    7578.0587  -2.855              0.004351
Zone_C         12143.9570    7983.3902   1.521              0.128377

Residual standard error: 34360 on 2037 degrees of freedom
Multiple R-squared:  0.8212,	Adjusted R-squared:  0.8204 
F-statistic:  1039 on 9 and 2037 DF,  p-value: < 0.00000000000000022
```

```
> summary(backward.lm)

Call:
lm(formula = SalePrice ~ OverallQual + YearBuilt + GrLivArea + 
    OC_SF + OQ_Rooms + Zone_FV + Zone_RH + Zone_RL + Zone_RM + 
    QualityIndex, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-360134  -19692   -2906   15984  245354 

Coefficients:
                 Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)  -888841.4853   73932.0996 -12.022 < 0.0000000000000002 ***
OverallQual    28079.0046    1496.1396  18.768 < 0.0000000000000002 ***
YearBuilt        427.8617      38.9559  10.983 < 0.0000000000000002 ***
GrLivArea         13.9455       3.5805   3.895             0.000101 ***
OC_SF              6.8936       0.3202  21.528 < 0.0000000000000002 ***
OQ_Rooms         803.8430     127.2204   6.319       0.000000000323 ***
Zone_FV       -21234.9522    8294.6344  -2.560             0.010536 *  
Zone_RH       -38705.8593   10375.4272  -3.731             0.000196 ***
Zone_RL       -17259.4542    7359.4197  -2.345             0.019111 *  
Zone_RM       -24656.8751    7404.4879  -3.330             0.000884 ***
QualityIndex   -1624.5192     173.9702  -9.338 < 0.0000000000000002 ***

Residual standard error: 34330 on 2036 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8207 
F-statistic: 937.2 on 10 and 2036 DF,  p-value: < 0.00000000000000022
```

```
> summary(stepwise.lm)

Call:
lm(formula = SalePrice ~ QualityIndex + OQ_Rooms + YearBuilt + 
    OC_SF + OverallQual + GrLivArea + Zone_RM + Zone_RH + Zone_C, 
    data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-359761  -19513   -3029   16086  245893 

Coefficients:
                 Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)  -885431.7901   74287.3402 -11.919 < 0.0000000000000002 ***
QualityIndex   -1643.9529     173.8329  -9.457 < 0.0000000000000002 ***
OQ_Rooms         823.5195     125.8949   6.541       0.000000000769 ***
YearBuilt        417.9492      38.7188  10.794 < 0.0000000000000002 ***
```

```
OC_SF              6.9209       0.3171  21.827 < 0.0000000000000002
OverallQual    27878.7571    1488.9207  18.724 < 0.0000000000000002
GrLivArea         13.6710       3.5560   3.844             0.000125
Zone_RM        -7650.6346    2368.4681  -3.230             0.001257
Zone_RH       -21633.1393    7578.0587  -2.855             0.004351
Zone_C         12143.9570    7983.3902   1.521             0.128377

Residual standard error: 34360 on 2037 degrees of freedom
Multiple R-squared:  0.8212,  Adjusted R-squared:  0.8204
F-statistic:  1039 on 9 and 2037 DF,  p-value: < 0.00000000000000022
> sort(vif(forward.lm),decreasing=TRUE)
 OverallQual     OQ_Rooms    GrLivArea QualityIndex        OC_SF    YearBuilt
Zone_RM       Zone_C      Zone_RH
    7.761173     7.122304     5.604607     4.557325     3.321212     2.416485
1.299086     1.069165     1.011023
> sort(vif(backward.lm),decreasing=TRUE)
     Zone_RL      Zone_RM  OverallQual     OQ_Rooms    GrLivArea      Zone_FV
QualityIndex        OC_SF    YearBuilt
   16.687389    12.717017     7.849119     7.284674     5.691086     5.657088
4.571808     3.392732     2.450072
     Zone_RH
    1.898232
> sort(vif(stepwise.lm),decreasing=TRUE)
 OverallQual     OQ_Rooms    GrLivArea QualityIndex        OC_SF    YearBuilt
Zone_RM       Zone_C      Zone_RH
    7.761173     7.122304     5.604607     4.557325     3.321212     2.416485
1.299086     1.069165     1.011023
```

Appendix 4: Prediction Grades

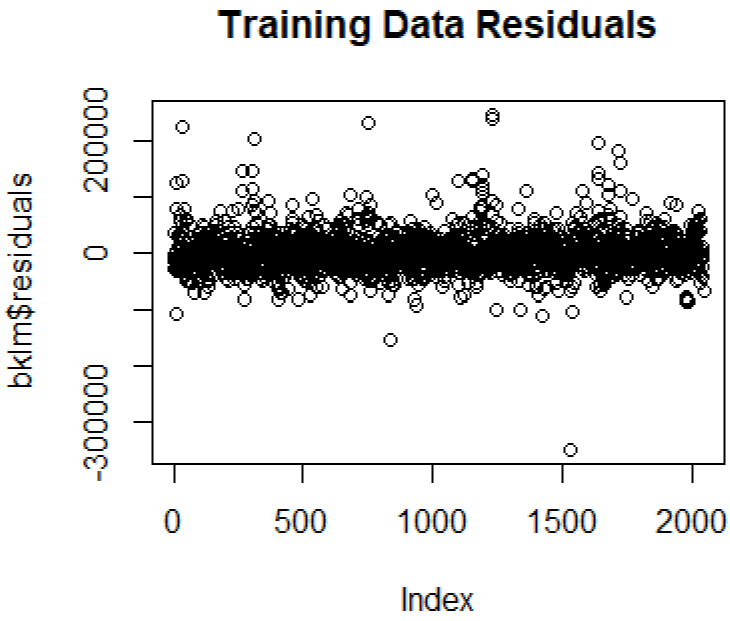
```
> forward.trainTable/sum(forward.trainTable)
forward.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.4792379            0.1704934            0.1954079            0.1
548608


> forward.testTable/sum(forward.testTable)
forward.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.5153584            0.1774744            0.1706485            0.1
365188


> backward.trainTable/sum(backward.trainTable)
backward.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.4758183            0.1709819            0.1968735            0.1
563263

> backward.testTable/sum(backward.testTable)
backward.testPredictionGrade
```

```
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.5164960            0.1729238            0.1740614            0.1
365188

> stepwise.trainTable/sum(stepwise.trainTable)
stepwise.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.4792379            0.1704934            0.1954079            0.1
548608

> stepwise.testTable/sum(stepwise.testTable)
stepwise.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.5153584            0.1774744            0.1706485            0.1
365188

> junk.trainTable/sum(junk.trainTable)
junk.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.4606742            0.1758671            0.1949194            0.1
685393

> junk.testTable/sum(junk.testTable)
junk.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.5017065            0.1786121            0.1717861            0.1
478953
```

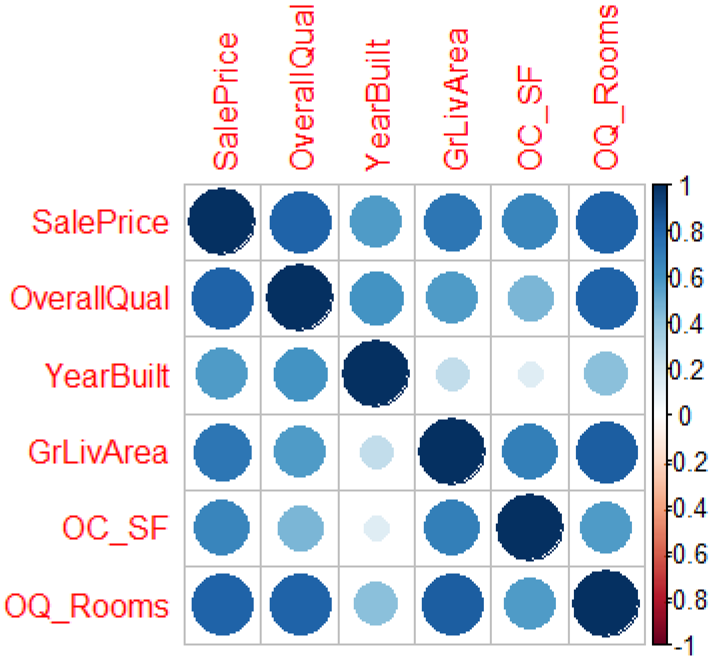
Appendix 5: Final Model Review

```
> summary(final)

Call:
lm(formula = SalePrice ~ OverallQual + YearBuilt + GrLivArea + 
    OC_SF + OQ_Rooms, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-350766  -19544   -3051   16287  245700 

Coefficients:
                 Estimate    Std. Error t value             Pr(>|t|)    
(Intercept) -1289924.3504    61459.3636 -20.988 < 0.0000000000000002
OverallQual    17871.1406     1124.5604  15.892 < 0.0000000000000002
YearBuilt        625.8537       32.3960  19.319 < 0.0000000000000002
GrLivArea         27.5187        3.3265   8.273 0.000000000000000234
OC_SF              5.0477        0.2499  20.201 < 0.0000000000000002
OQ_Rooms         790.0651      129.1274   6.118 0.000000001129336543

Residual standard error: 35310 on 2041 degrees of freedom
Multiple R-squared:  0.8108,	Adjusted R-squared:  0.8103 
```

```
F-statistic:  1749 on 5 and 2041 DF,  p-value: < 0.00000000000000022


> anova(final)
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq  F value                Pr(>F)
OverallQual    1 8617809767311 8617809767311 6912.665 < 0.00000000000000022
YearBuilt      1  171439269271  171439269271  137.518 < 0.00000000000000022
GrLivArea      1 1589843657794 1589843657794 1275.273 < 0.00000000000000022
OC_SF          1  476813641711  476813641711  382.470 < 0.00000000000000022
OQ_Rooms       1   46670274934   46670274934   37.436        0.000000001129
Residuals   2041 2544452757756    1246669651
```



```
> mse(final.test, test.df$SalePrice)
[1] 1664598806
> mae(final.test, test.df$SalePrice)
[1] 24219.33
```



```
> sort(vif(final),decreasing=TRUE)
   OQ_Rooms   GrLivArea OverallQual       OC_SF   YearBuilt
   7.095541    4.644459    4.192705    1.953238    1.602020
```

```
> final.trainTable/sum(final.trainTable)
final.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.4787494            0.1636541            0.2002931            0.1
573034
```

```
> final.testTable/sum(final.testTable)
final.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (
0.25+]
           0.5062571            0.1604096            0.1820250            0.1
513083
```