**Insights in Ames, Iowa: Real Estate**

MSDS 6371

Samuel Arellano, Eric Balke, Spencer Fogelman, Karen Somes

1. **Overview**

The selling prices of homes drive the real estate business. However, many factors influence a client’s purchasing decision and additionally the selling price. Aligning the client’s preferences with home attributes that also maximize selling price leads to satisfied buyers and optimal profitability. We will investigate what home characteristics are associated with selling price, so that Century 21 Ames can best position themselves in the market.

The Ames Housing Dataset compiled by Dean De Cock includes 79 explanatory variables to determine the corresponding selling price. There are 1,460 observations in the training data set and 1,459 in the testing data set. A full description of each variable can be found in the “data\_description” text file on Kaggle.com, and the variables included in the models can be found in the Appendix.

1. **Century 21 Analysis**

The square footage of the living area is a key component to a house, and potentially a major contributor to a client’s purchasing decision. The following analysis will assess the relationship between the area of living space in 100s of square feet in each of the neighborhoods that Century 21 deals in: North Ames, Edwards, and Brookside.

**Linear Regression Model**

To investigate the hypothesized relationship, a linear regression model was fitted on the variables of interest. Additionally, we will investigate whether the relationship differs between selling price and square footage based on neighborhood through interactions.





After fitting the model, all of the coefficients were found to be significantly different from zero at an alpha level of 0.05 (see Appendix).

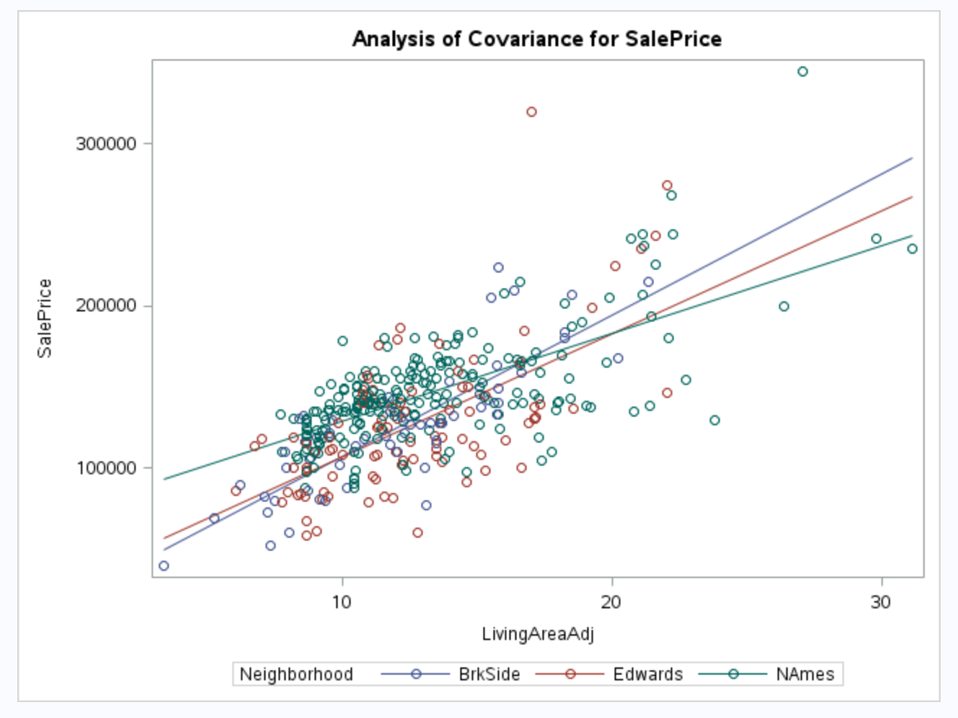




**Model Assumptions**

In order to accurately draw conclusions from the fitted model, several assumptions must be met: the residuals must be normally distributed with equal variance and no collinearity. Also, a linear relationship should exist between selling price and living space area.

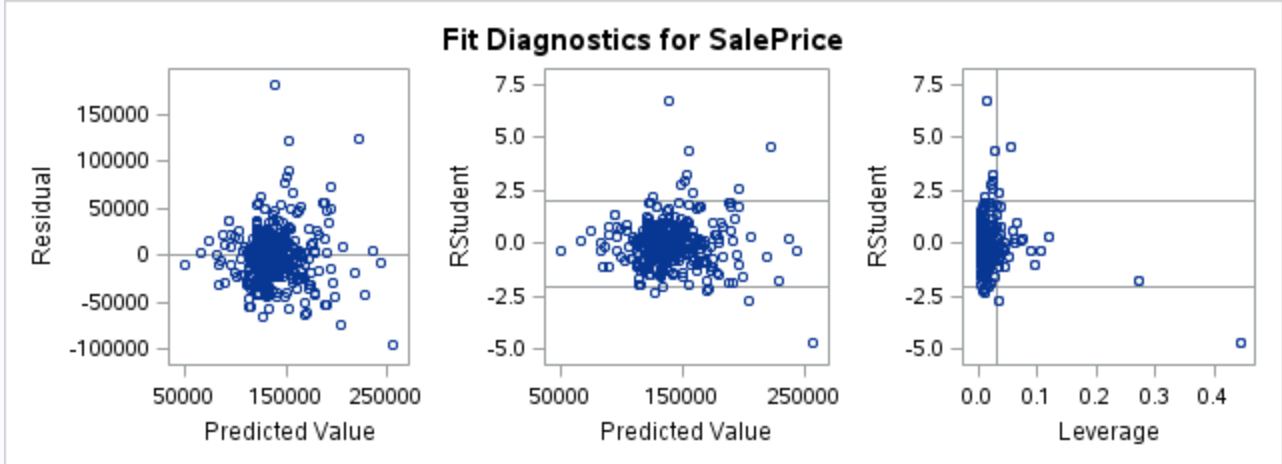
*Graph 1*



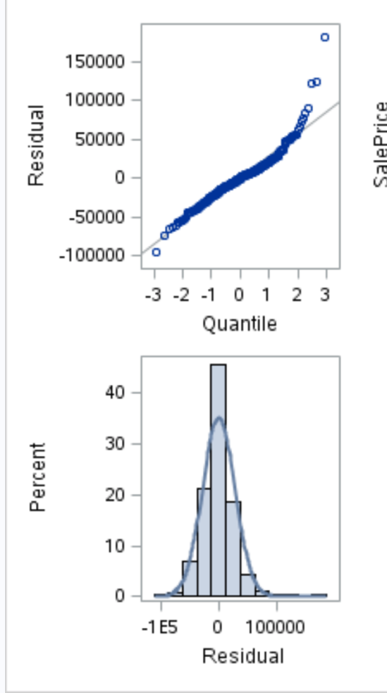
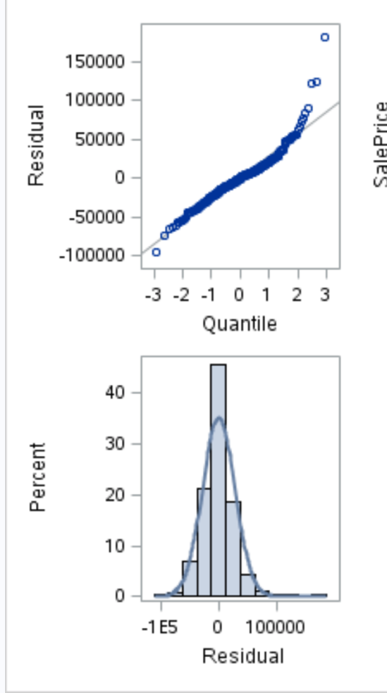
As shown in Graph 1, a linear relationship is a reasonable assumption for selling price and square footage. There are two points that appear potentially overly influential, however, shown by Cook’s D and the leverage values displayed in the Appendix. One outlier has a very high studentized residual and leverage, identified by Cook’s D. The other outlier has a less extreme studentized residual and is not identified by Cook’s D. However, the extremely large leverage values justify removing the two outliers to fit the model and perform the analysis.

The residuals appear to meet model assumptions as there is no apparent trend or unequal variances in the residual plot (Graph 2), and they appear normally distributed as reflected in the qq-plot (Graph 3).

*Graph 2*



*Graph 3*



**Model Performance**

Two key metrics can be used to assess how well the linear model fits the data given that the assumptions are met: Adjusted R2 and the internal Cross-validated PRESS value. The CV PRESS is 2.91 x 1011, and the R2 is 0.506. These values indicate that the predictive power of this model is not particularly strong, since the high cross-validated error and neighborhood and square footage of the living area only are associated with 50% of the variation in selling price.

**Parameter Estimates**

The fitted model reveals information about how square footage and neighborhood influence selling price, and the table of fitted values can be found in the Appendix.

For houses sold in the North Ames neighborhood, a house with zero square footage of living space has a selling value of $74,646. An increase in 100 square feet results in an increase in selling price of $5,431. A 95% confidence interval for this increase is A 95% CI for this increase is between $4579.28 and $6283.90.

Houses sold in Brookside have a different relationship where the starting value of the selling price for a house with 0 square footage is $19,942. With each increase of 100 square feet in living space the selling price increases by $8716. A 95% CI for this increase is (5431.57 + 1286.67, 5431.57 + 5282.67) = ($6718.24, $10714.24).

Likewise, for Edwards homes, the starting value of the selling price for a house with 0 square footage is $31,428. With each increase of 100 square feet in living space the selling price increases by $7597. A 95% CI for this increase is (5431.57 + 435.76, 5431.57+ 3896.36) = ($5867.33, $9327.93).

**Conclusions**

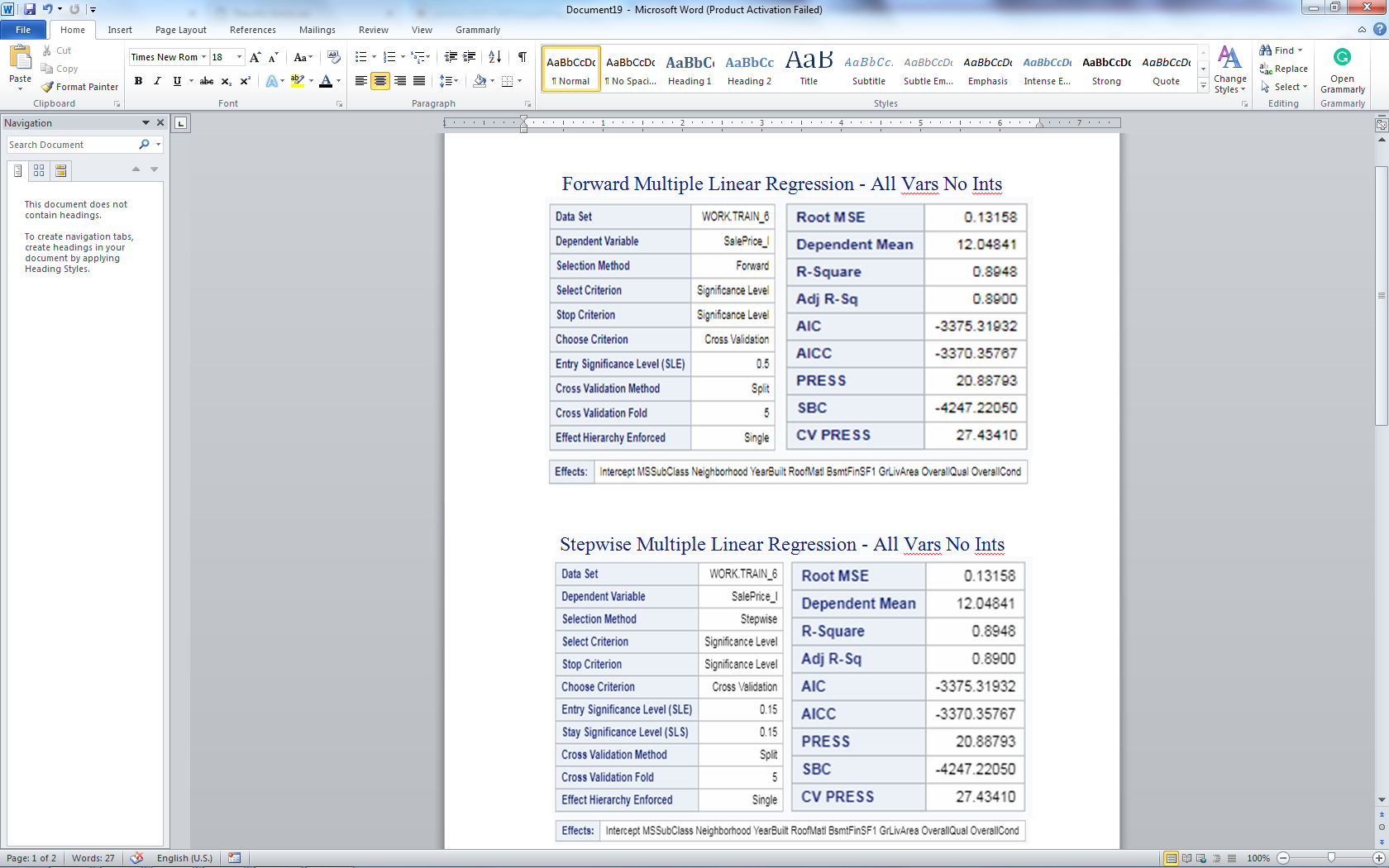
While the linear regression model provides useful information about the three variables of interest, the performance metrics indicate that other factors contribute to selling price. In the upcoming analysis, we will investigate other predictors to better understand how selling prices are influenced in Ames, Iowa.

1. **Predicting Selling Prices**

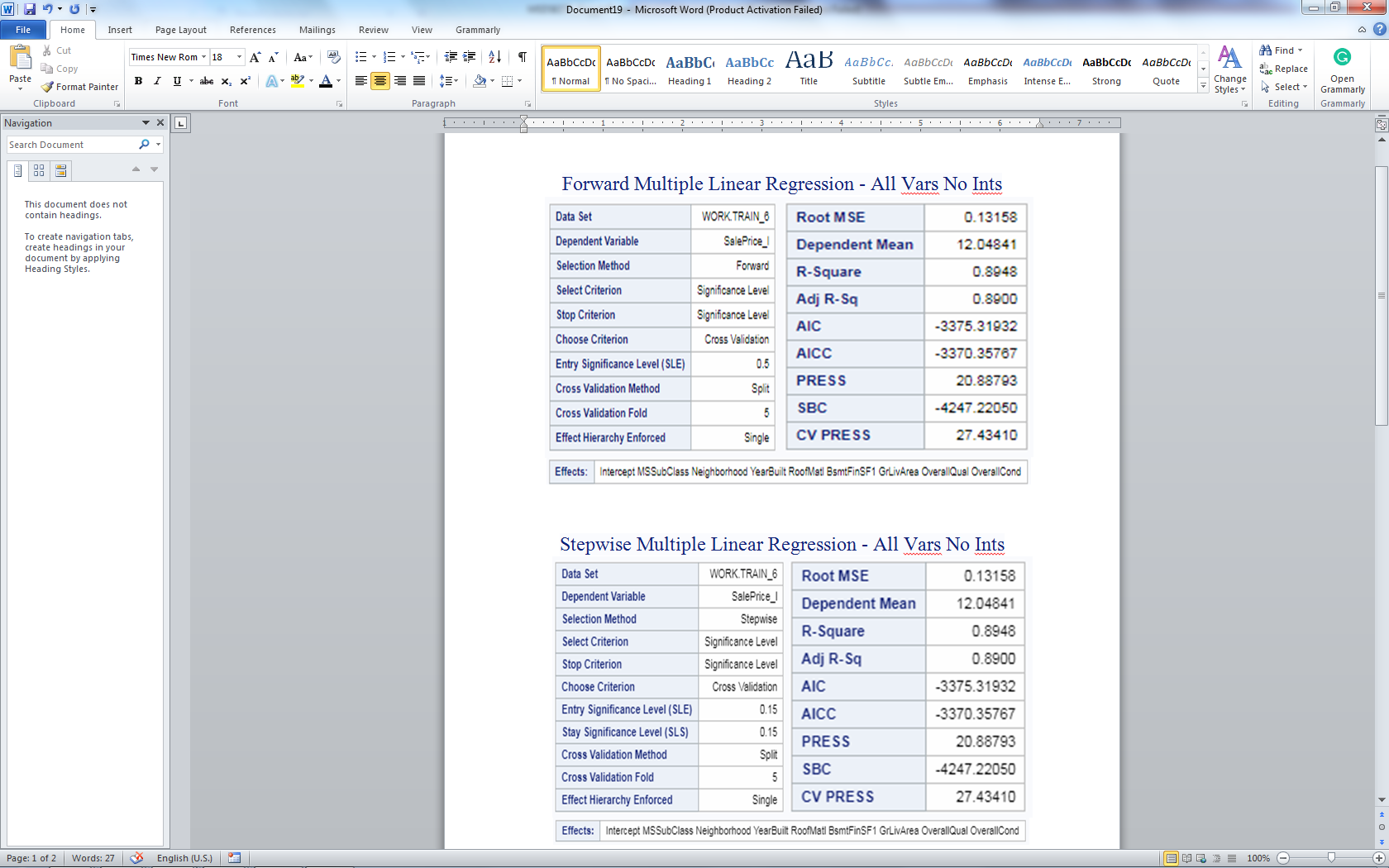
Determining the correlation between the predictive variables and the final selling price of a home allows us to go further into the analysis process and begin modeling. By creating statistical models and using regression techniques, we are able to create custom models that allow us to estimate the predicted selling price of a home based on readily available listing information. Being able to accurately predict the sales price of a home will allow Century 21 to better evaluate listing prices, estimate commissions, and allocate agent resources.

**Model Selection**

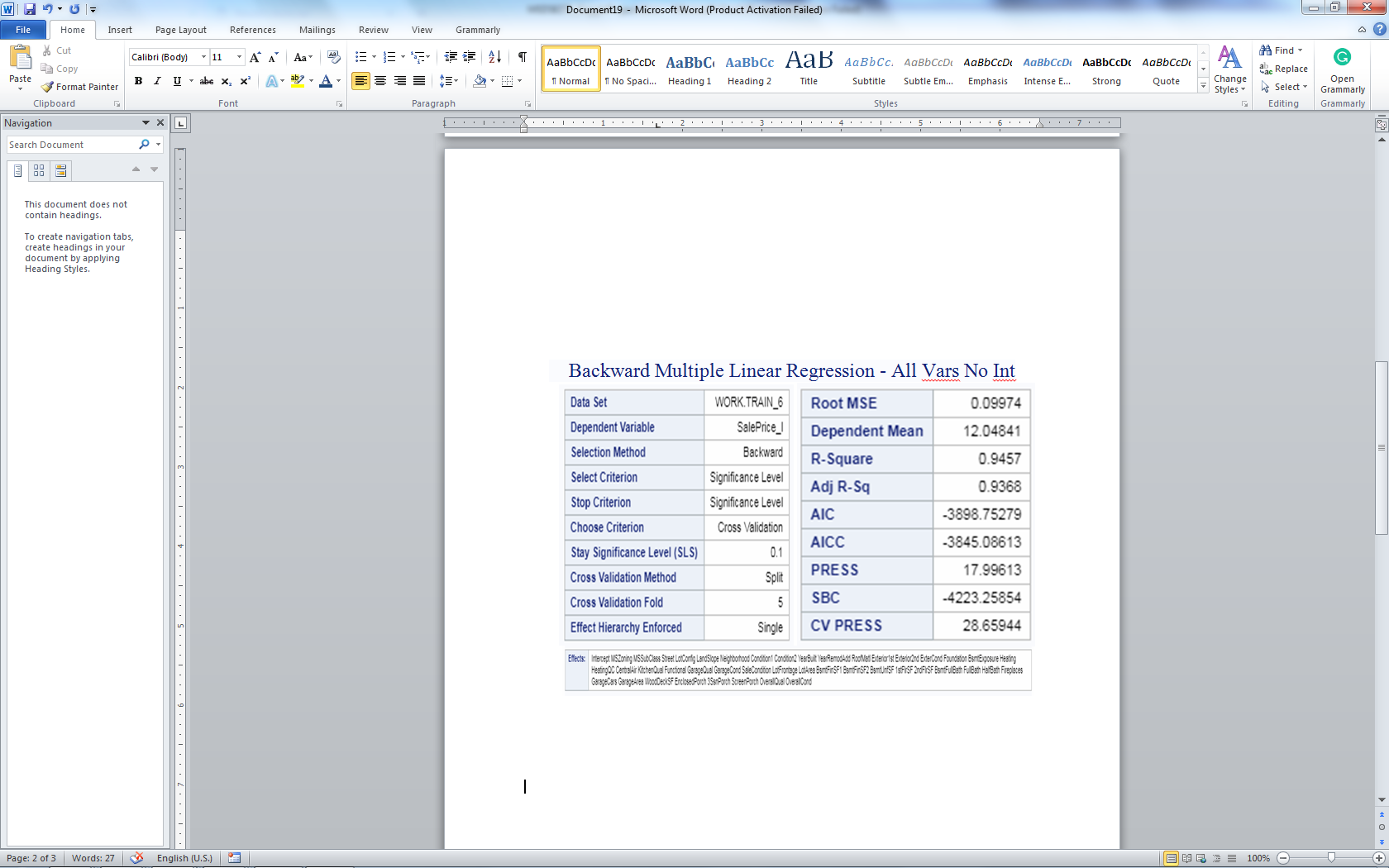
Based on the results of the exploratory data analysis, new variables were created for the log of the sales price and the log of the living area. For each model, the log of the sales price (SalePrice\_l) was used as the dependent variable. The first model selection process used was stepwise multiple linear regression using all explanatory variables and no interaction variables. Selection criterion were based on a significance level of .15 for both entry and stay significance. The step selection process resulted in a reduced model with 8 effect variables and a RMSE of .1316.



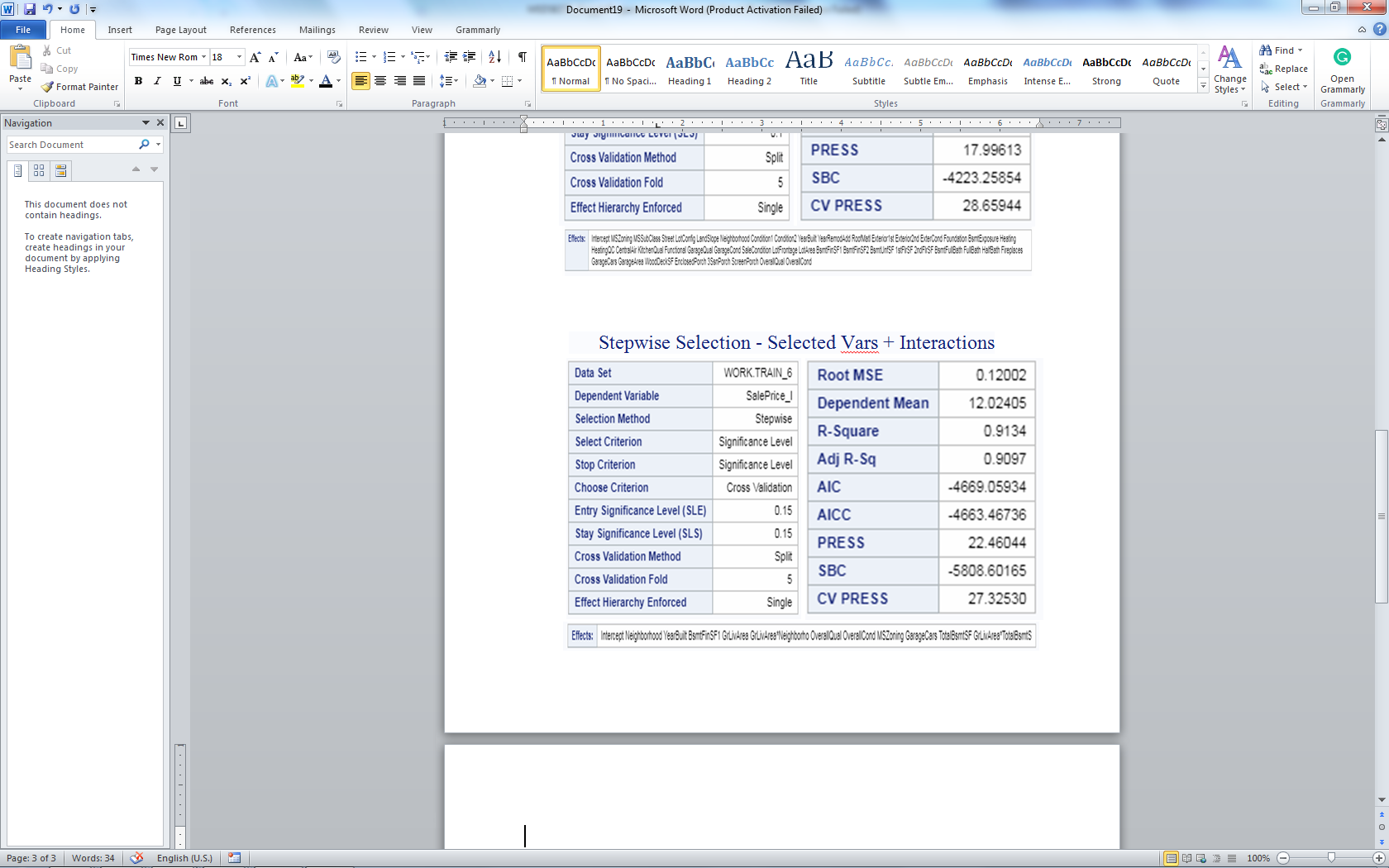
The next model was built using a forward multiple regression process checking all explanatory variables and no interaction variables. As with the stepwise regression, significance level was used as the selection criterion but the entry significance level was set to .5 and resulted in the same 8 variables and as a result, the same RMSE of .1316.



The third model was built using a backward selection process starting with all explanatory variables and no interaction variables. The stay significance for this model was .10 and expectedly the higher significance score led to a model with more included variables. The backward selection model had 33 variables and a RMSE of .0997.



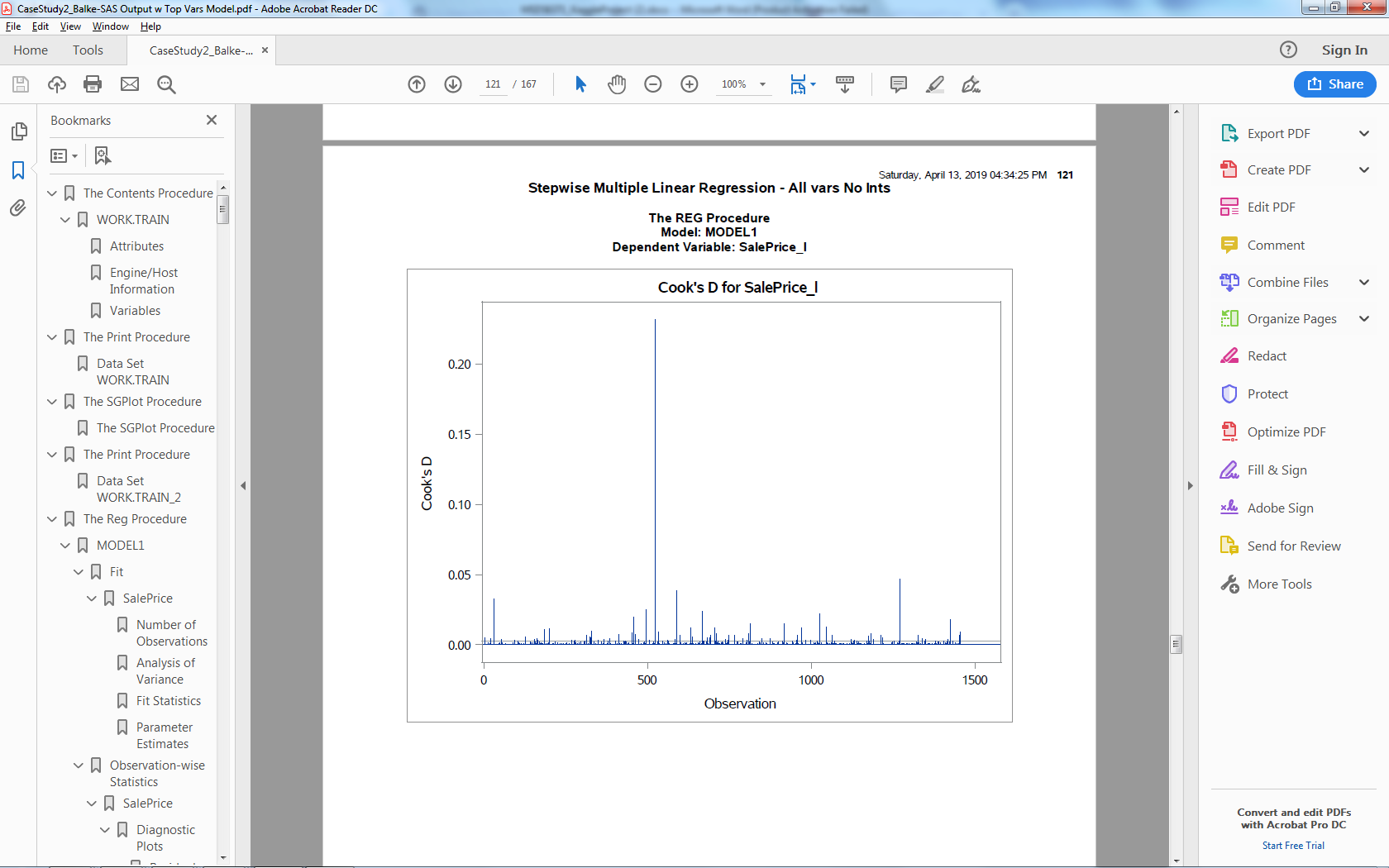
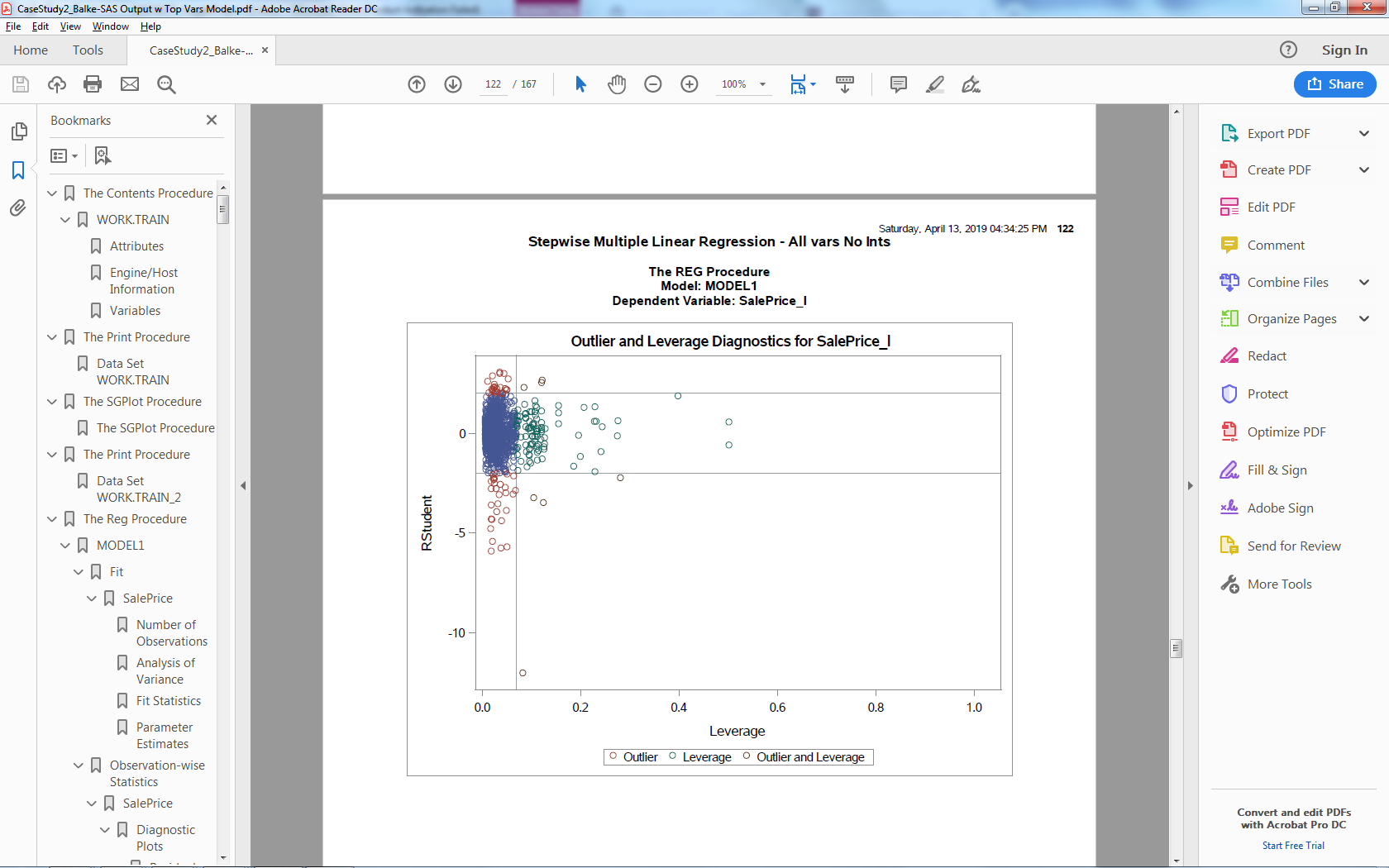
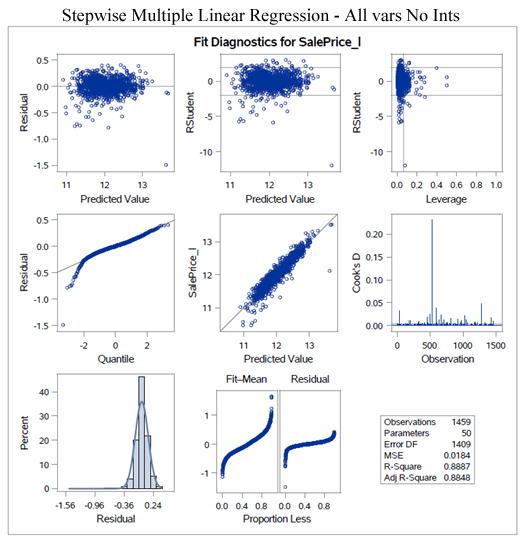
Finally a custom model was created using the stepwise process but allowing for interactions. Like the previous stepwise model, a .15 level of significance was used for both entry and stay. The resulting model had a total of 11 variables, 9 explanatory 2 interaction, and a RMSE of .1200.

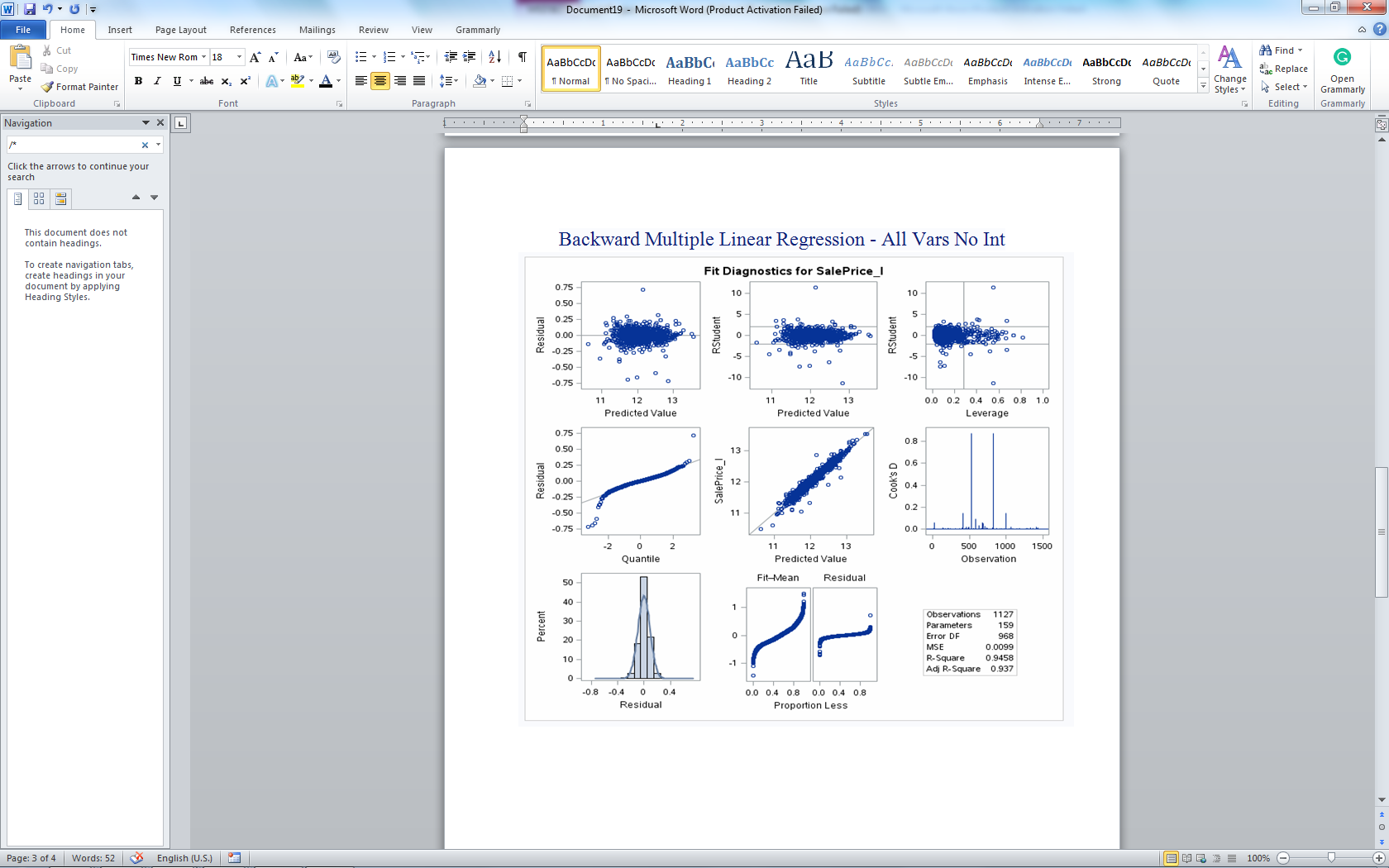


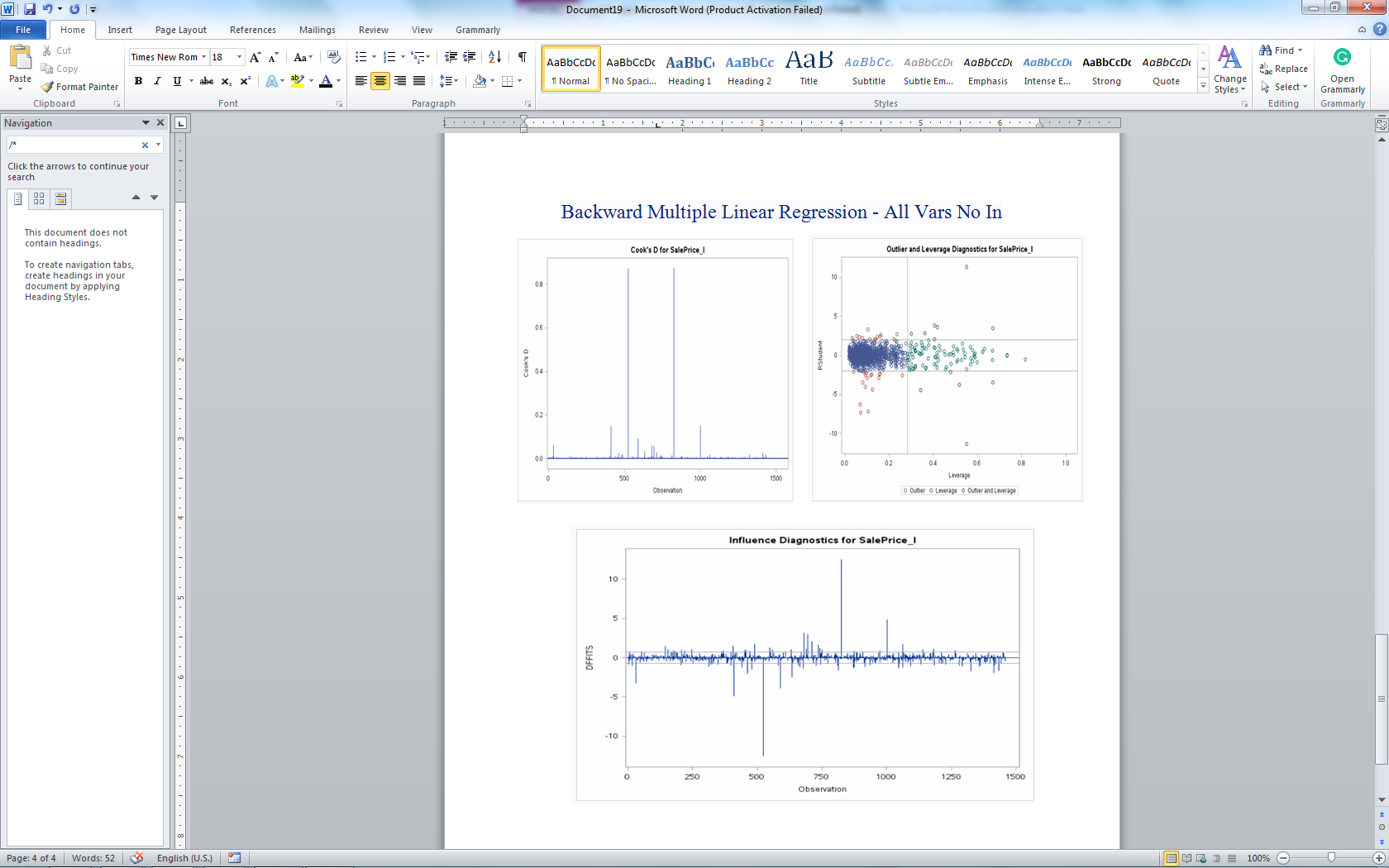
**Checking Assumptions**

Multilinear regression requires that certain assumptions be made about the data being used and that those assumptions be tested as part of the process. These assumptions are that there is a linear relationship (linearity) between the response variable and the explanatory variables, that there is a normal distribution of residuals (normality), and that independent variables are not highly correlated with each other (no multicollinearity). Linearity was checked by reviewing scatterplots of the residuals. Normality was verified by reviewing residual scatterplots and by analyzing the impact of outlier data points. Multicollinearity was checked through Variance Inflation Factor (VIF) values and using model selection processes that reduced the effects of collinearity.

Regression plot data for simple stepwise and forward selection were the same as they both yielded the same models. Residual scatterplots indicated no reason to suggest that the data was not normally distributes and indicated strong linearity between the response and explanatory variables. Cook’s D and Outlier Leverage Diagnostics indicated the presence of outliers but the low Cook’s score of the outlier point and the general distribution of outlying points within the minimal leverage zone suggests that their effect on the regression is negligible. It should be noted that the fit tests do not result in a straight line because the dependent variable was logged.

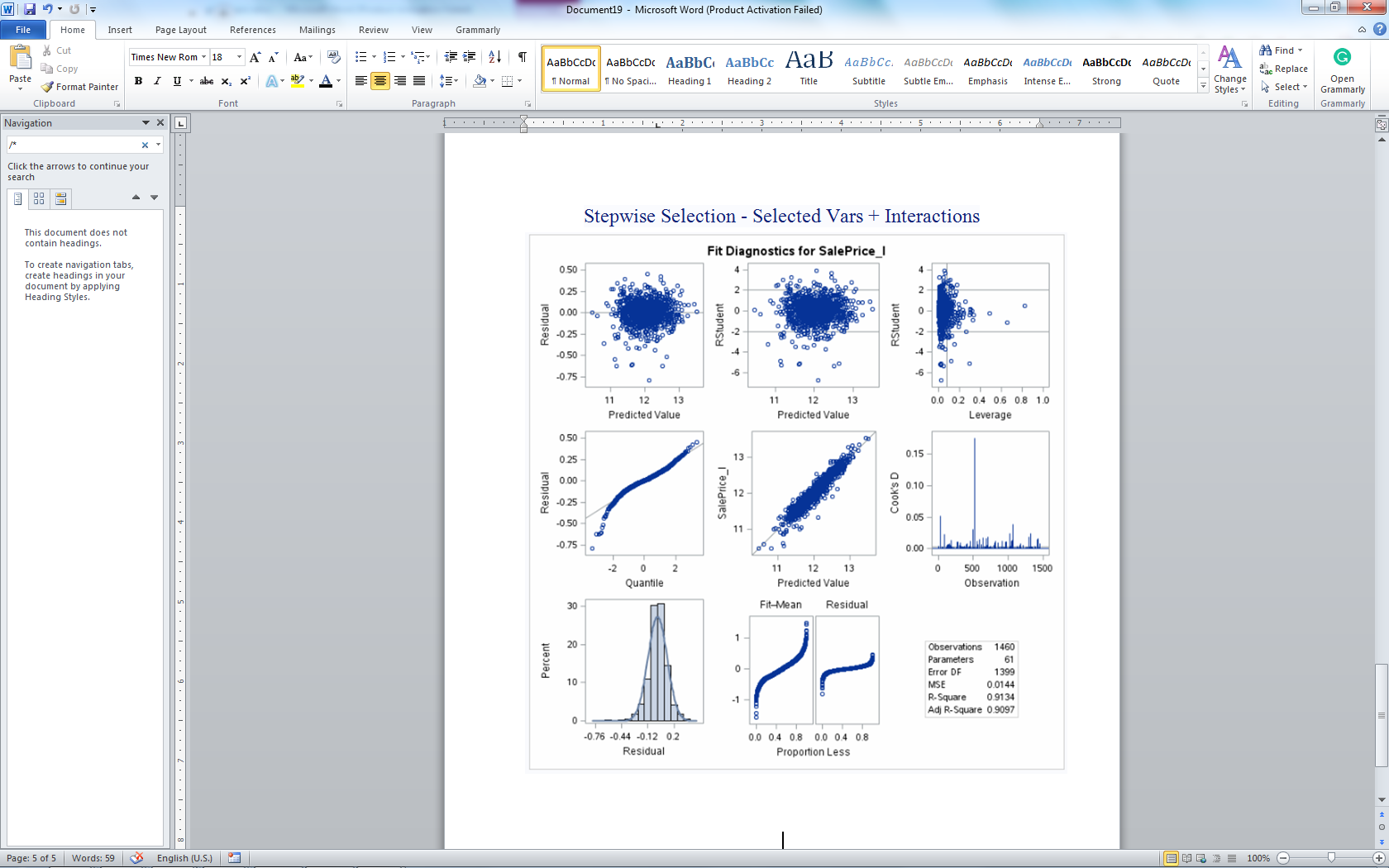
****

As with the stepwise and forward selection, residual scatterplots for the backward selection model indicated linearity and normality, however the addition of more variables also resulted in the introduction of more outliers.



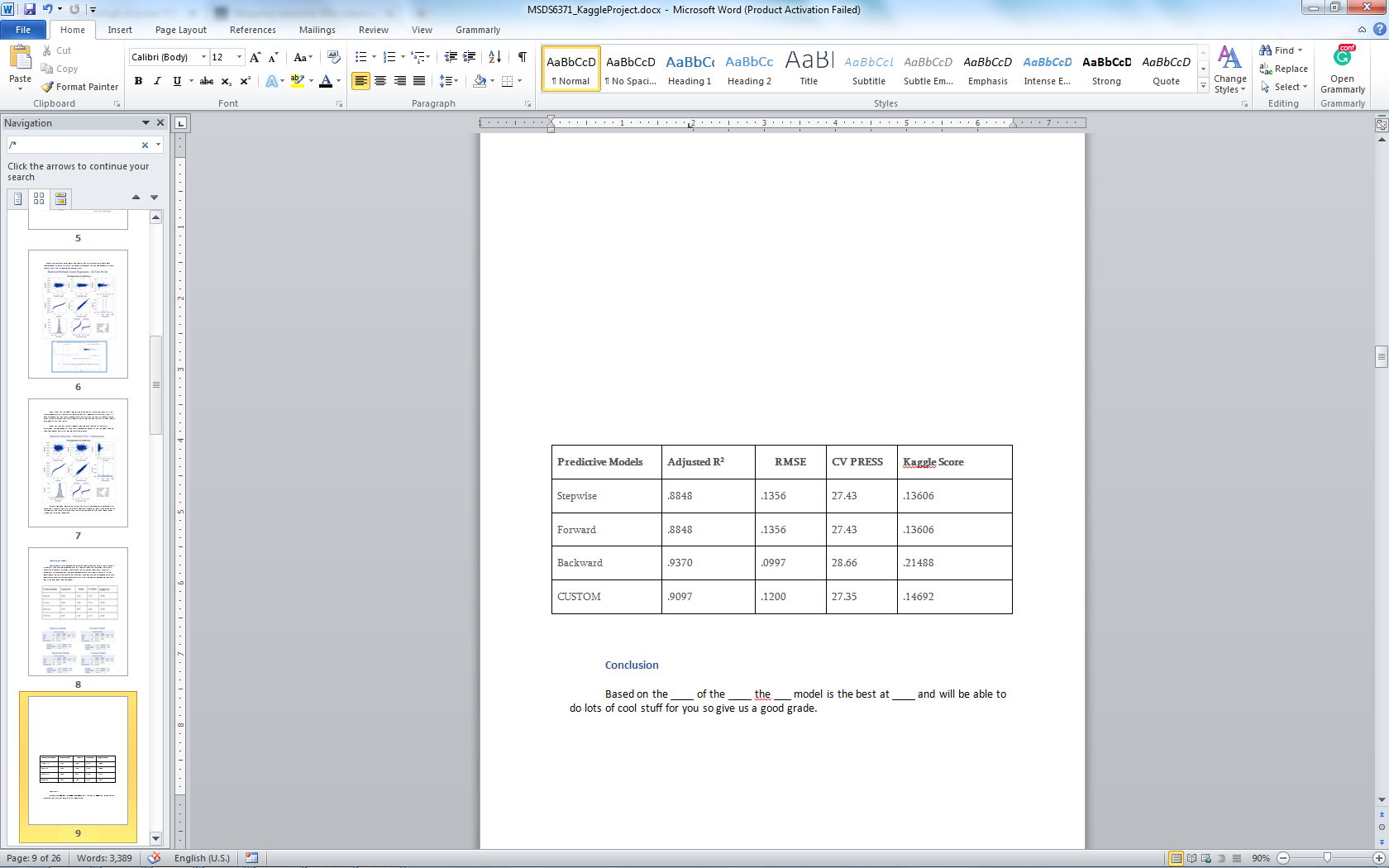
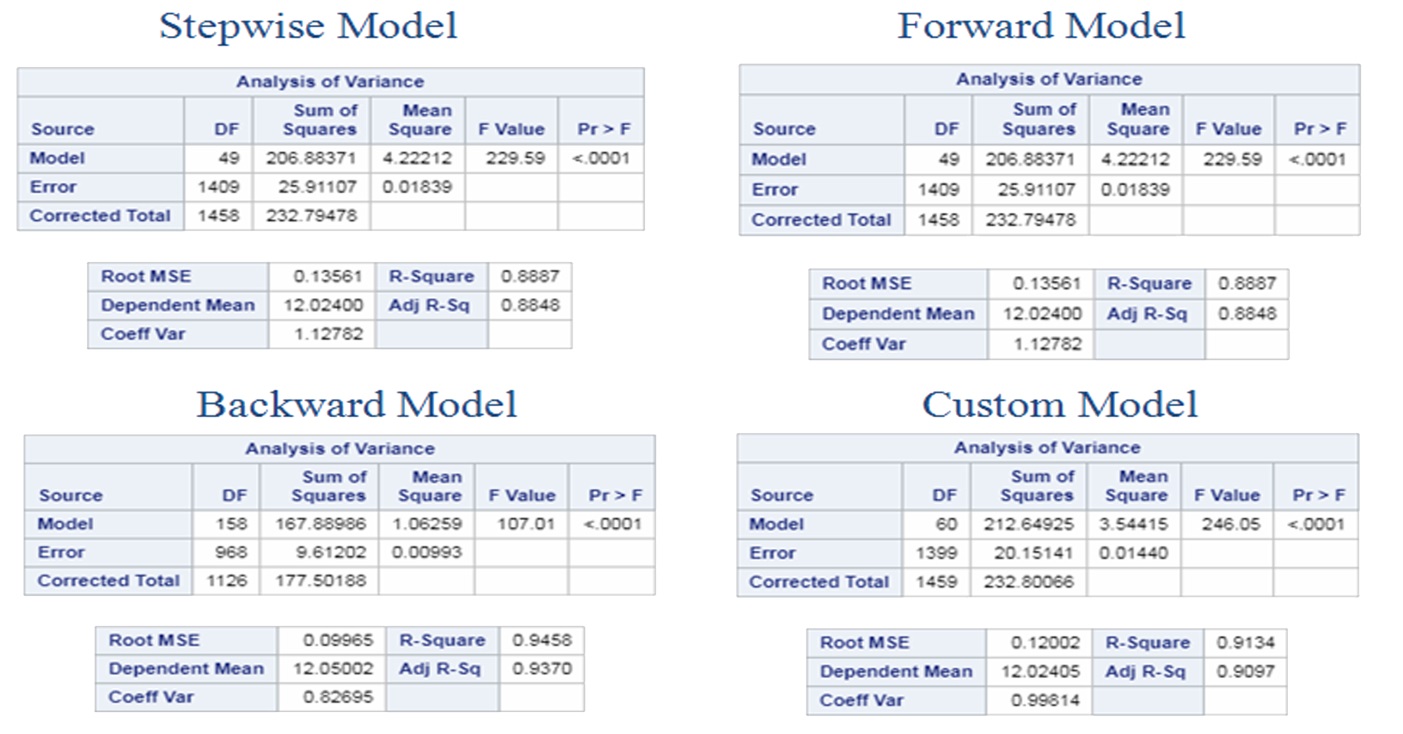
Visual analysis of the Cook’s D graph indicated outliers with values close to 1 and therefore warranted a closer evaluation of the outliers. However, after further analysis it was determined that due to the symmetry of the outliers and their overall impact and the lack of reason to believe they values where errors, it was decided that the variables remain in the model as valid variables.

Analysis of residuals for the stepwise selection with selected variables and interactions (custom) model satisfied the assumption of normality and the Cook’s D value and disbursement of outliers was well within tolerances.



As such all models were tested for their accuracy in determining an unknown sales price given all other required variables for the model. Each model was trained based on the available data and in the case of missing sales prices the median sales price was used for training and testing of the models.

**Comparison of Models**

The stepwise and forward models were the simplest models each using the same 8 variables as such they performed equally in all metrics. They had the lowest adjusted R2 value, highest RMSE, and lowest Kaggle score, indicating they were more accurate at predicting the actual sale price than the other models. The backward model had the highest adjusted R2 value, lowest RMSE, highest CV PRESS, and had the worst Kaggle Score. The custom model had the lowest CV PRESS value and performed worse than the stepwise/forward model, but better than the backward selection model.

**Conclusion**

Based on our analysis the best model to utilize is the stepwise selectin model as it is not only the simplest model but also scored better than the other models when it came to predicting the sales price of a home.

**Appendix**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/  
/\* Century 21 Analysis \*/  
/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

options mlogic symbolgen;  
  
data train;  
 %let \_EFIERR\_ = 0;  
  
 /\* set the ERROR detection macro variable \*/  
 infile '/home/ spencer.fogelman/CaseStudy2/train.csv' delimiter=','

MISSOVER DSD   
 lrecl=32767 firstobs=2;   
 informat Id best32.;  
 informat MSSubClass $3.;  
 informat MSZoning $2.;  
 informat LotFrontage best32.;  
 informat LotArea best32.;  
 informat Street $4.;  
 informat Alley $2.;  
 informat LotShape $3.;  
 informat LandContour $3.;  
 informat Utilities $6.;  
 informat LotConfig $7.;  
 informat LandSlope $3.;  
 informat Neighborhood $7.;  
 informat Condition1 $6.;  
 informat Condition2 $6.;  
 informat BldgType $6.;  
 informat HouseStyle $6.;  
 informat OverallQual best32.;  
 informat OverallCond best32.;  
 informat YearBuilt best32.;  
 informat YearRemodAdd best32.;  
 informat RoofStyle $5.;  
 informat RoofMatl $7.;  
 informat Exterior1st $7.;  
 informat Exterior2nd $7.;  
 informat MasVnrType $7.;  
 informat MasVnrArea best32.;  
 informat ExterQual $2.;  
 informat ExterCond $2.;  
 informat Foundation $6.;  
 informat BsmtQual $2.;  
 informat BsmtCond $2.;  
 informat BsmtExposure $2.;  
 informat BsmtFinType1 $3.;  
 informat BsmtFinSF1 best32.;  
 informat BsmtFinType2 $3.;  
 informat BsmtFinSF2 best32.;  
 informat BsmtUnfSF best32.;  
 informat TotalBsmtSF best32.;  
 informat Heating $4.;  
 informat HeatingQC $2.;  
 informat CentralAir $1.;  
 informat Electrical $5.;  
 informat "1stFlrSF"N best32.;  
 informat "2ndFlrSF"N best32.;  
 informat LowQualFinSF best32.;  
 informat GrLivArea best32.;  
 informat BsmtFullBath best32.;  
 informat BsmtHalfBath best32.;  
 informat FullBath best32.;  
 informat HalfBath best32.;  
 informat BedroomAbvGr best32.;  
 informat KitchenAbvGr best32.;  
 informat KitchenQual $2.;  
 informat TotRmsAbvGrd best32.;  
 informat Functional $4.;  
 informat Fireplaces best32.;  
 informat FireplaceQu $2.;  
 informat GarageType $7.;  
 informat GarageYrBlt best32.;  
 informat GarageFinish $3.;  
 informat GarageCars best32.;  
 informat GarageArea best32.;  
 informat GarageQual $2.;  
 informat GarageCond $2.;  
 informat PavedDrive $1.;  
 informat WoodDeckSF best32.;  
 informat OpenPorchSF best32.;  
 informat EnclosedPorch best32.;  
 informat "3SsnPorch"N best32.;  
 informat ScreenPorch best32.;  
 informat PoolArea best32.;  
 informat PoolQC $2.;  
 informat Fence $5.;  
 informat MiscFeature $4.;  
 informat MiscVal best32.;  
 informat MoSold best32.;  
 informat YrSold best32.;  
 informat SaleType $3.;  
 informat SaleCondition $7.;  
 informat SalePrice best32.;  
 format Id best12.;  
 format MSSubClass $3.;  
 format MSZoning $2.;  
 format LotFrontage best12.;  
 format LotArea best12.;  
 format Street $4.;  
 format Alley $2.;  
 format LotShape $3.;  
 format LandContour $3.;  
 format Utilities $6.;  
 format LotConfig $7.;  
 format LandSlope $3.;  
 format Neighborhood $7.;  
 format Condition1 $6.;  
 format Condition2 $6.;  
 format BldgType $6.;  
 format HouseStyle $6.;  
 format OverallQual best12.;  
 format OverallCond best12.;  
 format YearBuilt best12.;  
 format YearRemodAdd best12.;  
 format RoofStyle $5.;  
 format RoofMatl $7.;  
 format Exterior1st $7.;  
 format Exterior2nd $7.;  
 format MasVnrType $7.;  
 format MasVnrArea best12.;  
 format ExterQual $2.;  
 format ExterCond $2.;  
 format Foundation $6.;  
 format BsmtQual $2.;  
 format BsmtCond $2.;  
 format BsmtExposure $2.;  
 format BsmtFinType1 $3.;  
 format BsmtFinSF1 best12.;  
 format BsmtFinType2 $3.;  
 format BsmtFinSF2 best12.;  
 format BsmtUnfSF best12.;  
 format TotalBsmtSF best12.;  
 format Heating $4.;  
 format HeatingQC $2.;  
 format CentralAir $1.;  
 format Electrical $5.;  
 format "1stFlrSF"N best12.;  
 format "2ndFlrSF"N best12.;  
 format LowQualFinSF best12.;  
 format GrLivArea best12.;  
 format BsmtFullBath best12.;  
 format BsmtHalfBath best12.;  
 format FullBath best12.;  
 format HalfBath best12.;  
 format BedroomAbvGr best12.;  
 format KitchenAbvGr best12.;  
 format KitchenQual $2.;  
 format TotRmsAbvGrd best12.;  
 format Functional $4.;  
 format Fireplaces best12.;  
 format FireplaceQu $2.;  
 format GarageType $7.;  
 format GarageYrBlt best12.;  
 format GarageFinish $3.;  
 format GarageCars best12.;  
 format GarageArea best12.;  
 format GarageQual $2.;  
 format GarageCond $2.;  
 format PavedDrive $1.;  
 format WoodDeckSF best12.;  
 format OpenPorchSF best12.;  
 format EnclosedPorch best12.;  
 format "3SsnPorch"N best12.;  
 format ScreenPorch best12.;  
 format PoolArea best12.;  
 format PoolQC $2.;  
 format Fence $5.;  
 format MiscFeature $4.;  
 format MiscVal best12.;  
 format MoSold best12.;  
 format YrSold best12.;  
 format SaleType $3.;  
 format SaleCondition $7.;  
 format SalePrice best12.;  
 input Id MSSubClass $  
 MSZoning $  
 LotFrontage LotArea Street $  
 Alley $  
 LotShape $  
 LandContour $  
 Utilities $  
 LotConfig $  
 LandSlope $  
 Neighborhood $  
 Condition1 $  
 Condition2 $  
 BldgType $  
 HouseStyle $  
 OverallQual OverallCond YearBuilt YearRemodAdd   
 RoofStyle $  
 RoofMatl $  
 Exterior1st $  
 Exterior2nd $  
 MasVnrType $  
 MasVnrArea ExterQual $  
 ExterCond $  
 Foundation $  
 BsmtQual $  
 BsmtCond $  
 BsmtExposure $  
 BsmtFinType1 $  
 BsmtFinSF1 BsmtFinType2 $  
 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating $  
 HeatingQC $  
 CentralAir $  
 Electrical $  
 "1stFlrSF"N "2ndFlrSF"N LowQualFinSF GrLivArea   
 BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr

KitchenAbvGr   
 KitchenQual $  
 TotRmsAbvGrd Functional $  
 Fireplaces FireplaceQu $  
 GarageType $  
 GarageYrBlt GarageFinish $  
 GarageCars GarageArea GarageQual $  
 GarageCond $  
 PavedDrive $  
 WoodDeckSF OpenPorchSF EnclosedPorch "3SsnPorch"N   
 ScreenPorch PoolArea PoolQC $  
 Fence $  
 MiscFeature $  
 MiscVal MoSold YrSold SaleType $  
 SaleCondition $  
 SalePrice;  
  
 if \_ERROR\_ then  
 call symputx('\_EFIERR\_', 1);  
  
 /\* set ERROR detection macro variable \*/  
run;  
  
proc contents data=train;  
run;  
  
proc print data=train (obs=10);  
run;  
  
data train\_1;  
 set train;  
 where Neighborhood in ("NAmes", "Edwards", "BrkSide");  
run;  
  
data train\_2;  
 set train\_1;  
 N\_Dummy=0;  
 E\_Dummy=0;  
  
 if Neighborhood="NAmes" then  
 N\_Dummy=1;  
 else if Neighborhood="Edwards" then  
 E\_Dummy=1;  
 int\_e=E\_Dummy \* GrLIvArea;  
 int\_n=N\_Dummy \* GrLIvArea;  
run;  
  
title "Scatterplot of original data";  
  
proc sgplot data=train\_2;  
 scatter x=GrLIvArea y=SalePrice / datalabel=Id   
 markerattrs=(symbol=CircleFilled) group=Neighborhood;  
run;  
  
title;  
  
proc print data=train\_2 (obs=10);  
run;  
  
ods graphics on;  
title   
 "Linear-Linear MLR with Dummy and Interactions with Influential Point 1299";  
  
proc reg data=train\_2 plots=all;  
 model SalePrice=GrLIvArea E\_Dummy N\_Dummy int\_e int\_n;  
 ID id;  
 run;  
quit;  
  
title;  
ods graphics off;  
  
data train\_3;  
 set train\_2;  
 where Id ne 1299;  
run;  
  
ods graphics on;  
title   
 "Linear-Linear MLR with Dummy and Interactions w/o Influential Point 1299";  
  
proc reg data=train\_3 plots=all;  
 model SalePrice=GrLIvArea E\_Dummy N\_Dummy int\_e int\_n;  
 ID id;  
 run;  
quit;  
  
title;  
ods graphics off;  
  
data train\_4;  
 set train\_2;  
 SalePrice\_l=log(SalePrice);  
 GrLIvArea\_l=log(GrLIvArea);  
 int\_e\_l=E\_Dummy \* GrLIvArea\_l;  
 int\_n\_l=N\_Dummy \* GrLIvArea\_l;  
run;  
  
ods graphics on;  
title "Log-Linear MLR with Dummy and Interactions";  
  
proc reg data=train\_4 plots=all;  
 model SalePrice\_l=GrLIvArea E\_Dummy N\_Dummy int\_e int\_n;  
 ID id;  
 run;  
quit;  
  
title;  
ods graphics off;  
ods graphics on;  
title "Linear-Log MLR with Dummy and Interactions";  
  
proc reg data=train\_4 plots=all;  
 model SalePrice=GrLIvArea\_l E\_Dummy N\_Dummy int\_e\_l int\_n\_l;  
 ID id;  
 run;  
quit;  
  
title;  
ods graphics off;  
ods graphics on;  
title "Log-Log MLR with Dummy and Interactions";  
  
proc reg data=train\_4 plots=all;  
 model SalePrice\_l=GrLIvArea\_l E\_Dummy N\_Dummy int\_e\_l int\_n\_l;  
 ID id;  
 run;  
quit;  
  
title;  
ods graphics off;  
  
proc means data=train\_4;  
 var GrLIvArea\_l E\_Dummy N\_Dummy;  
run;  
  
data train\_5;  
 set train\_4;  
 centE=(GrLIvArea\_l - 7.1193132)\*(E\_Dummy - 0.2610966);  
 centN=(GrLIvArea\_l - 7.1193132)\*(N\_Dummy - 0.5874674);  
run;  
  
ods graphics on;  
title "Log-Log MLR with Dummy and Centered Interactions";  
  
proc reg data=train\_5 plots=all;  
 model SalePrice\_l=GrLIvArea\_l E\_Dummy N\_Dummy centE centN;  
 ID id;  
 run;  
quit;  
  
title;  
ods graphics off;  
  
/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/  
/\* Predicting Selling Prices \*/  
/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Import the test data \*/  
data test;  
 %let \_EFIERR\_ = 0;  
  
 /\* set the ERROR detection macro variable \*/  
 infile '/home/eric.balke/sasuser.v94/CaseStudy2/test.csv' delimiter=','

MISSOVER DSD   
 lrecl=32767 firstobs=2;   
 informat Id best32.;  
 informat MSSubClass $3.;  
 informat MSZoning $2.;  
 informat LotFrontage best32.;  
 informat LotArea best32.;  
 informat Street $4.;  
 informat Alley $2.;  
 informat LotShape $3.;  
 informat LandContour $3.;  
 informat Utilities $6.;  
 informat LotConfig $7.;  
 informat LandSlope $3.;  
 informat Neighborhood $7.;  
 informat Condition1 $6.;  
 informat Condition2 $6.;  
 informat BldgType $6.;  
 informat HouseStyle $6.;  
 informat OverallQual best32.;  
 informat OverallCond best32.;  
 informat YearBuilt best32.;  
 informat YearRemodAdd best32.;  
 informat RoofStyle $5.;  
 informat RoofMatl $7.;  
 informat Exterior1st $7.;  
 informat Exterior2nd $7.;  
 informat MasVnrType $7.;  
 informat MasVnrArea best32.;  
 informat ExterQual $2.;  
 informat ExterCond $2.;  
 informat Foundation $6.;  
 informat BsmtQual $2.;  
 informat BsmtCond $2.;  
 informat BsmtExposure $2.;  
 informat BsmtFinType1 $3.;  
 informat BsmtFinSF1 best32.;  
 informat BsmtFinType2 $3.;  
 informat BsmtFinSF2 best32.;  
 informat BsmtUnfSF best32.;  
 informat TotalBsmtSF best32.;  
 informat Heating $4.;  
 informat HeatingQC $2.;  
 informat CentralAir $1.;  
 informat Electrical $5.;  
 informat "1stFlrSF"N best32.;  
 informat "2ndFlrSF"N best32.;  
 informat LowQualFinSF best32.;  
 informat GrLivArea best32.;  
 informat BsmtFullBath best32.;  
 informat BsmtHalfBath best32.;  
 informat FullBath best32.;  
 informat HalfBath best32.;  
 informat BedroomAbvGr best32.;  
 informat KitchenAbvGr best32.;  
 informat KitchenQual $2.;  
 informat TotRmsAbvGrd best32.;  
 informat Functional $4.;  
 informat Fireplaces best32.;  
 informat FireplaceQu $2.;  
 informat GarageType $7.;  
 informat GarageYrBlt best32.;  
 informat GarageFinish $3.;  
 informat GarageCars best32.;  
 informat GarageArea best32.;  
 informat GarageQual $2.;  
 informat GarageCond $2.;  
 informat PavedDrive $1.;  
 informat WoodDeckSF best32.;  
 informat OpenPorchSF best32.;  
 informat EnclosedPorch best32.;  
 informat "3SsnPorch"N best32.;  
 informat ScreenPorch best32.;  
 informat PoolArea best32.;  
 informat PoolQC $2.;  
 informat Fence $5.;  
 informat MiscFeature $4.;  
 informat MiscVal best32.;  
 informat MoSold best32.;  
 informat YrSold best32.;  
 informat SaleType $3.;  
 informat SaleCondition $7.;  
 format Id best12.;  
 format MSSubClass $3.;  
 format MSZoning $2.;  
 format LotFrontage best12.;  
 format LotArea best12.;  
 format Street $4.;  
 format Alley $2.;  
 format LotShape $3.;  
 format LandContour $3.;  
 format Utilities $6.;  
 format LotConfig $7.;  
 format LandSlope $3.;  
 format Neighborhood $7.;  
 format Condition1 $6.;  
 format Condition2 $6.;  
 format BldgType $6.;  
 format HouseStyle $6.;  
 format OverallQual best12.;  
 format OverallCond best12.;  
 format YearBuilt best12.;  
 format YearRemodAdd best12.;  
 format RoofStyle $5.;  
 format RoofMatl $7.;  
 format Exterior1st $7.;  
 format Exterior2nd $7.;  
 format MasVnrType $7.;  
 format MasVnrArea best12.;  
 format ExterQual $2.;  
 format ExterCond $2.;  
 format Foundation $6.;  
 format BsmtQual $2.;  
 format BsmtCond $2.;  
 format BsmtExposure $2.;  
 format BsmtFinType1 $3.;  
 format BsmtFinSF1 best12.;  
 format BsmtFinType2 $3.;  
 format BsmtFinSF2 best12.;  
 format BsmtUnfSF best12.;  
 format TotalBsmtSF best12.;  
 format Heating $4.;  
 format HeatingQC $2.;  
 format CentralAir $1.;  
 format Electrical $5.;  
 format "1stFlrSF"N best12.;  
 format "2ndFlrSF"N best12.;  
 format LowQualFinSF best12.;  
 format GrLivArea best12.;  
 format BsmtFullBath best12.;  
 format BsmtHalfBath best12.;  
 format FullBath best12.;  
 format HalfBath best12.;  
 format BedroomAbvGr best12.;  
 format KitchenAbvGr best12.;  
 format KitchenQual $2.;  
 format TotRmsAbvGrd best12.;  
 format Functional $4.;  
 format Fireplaces best12.;  
 format FireplaceQu $2.;  
 format GarageType $7.;  
 format GarageYrBlt best12.;  
 format GarageFinish $3.;  
 format GarageCars best12.;  
 format GarageArea best12.;  
 format GarageQual $2.;  
 format GarageCond $2.;  
 format PavedDrive $1.;  
 format WoodDeckSF best12.;  
 format OpenPorchSF best12.;  
 format EnclosedPorch best12.;  
 format "3SsnPorch"N best12.;  
 format ScreenPorch best12.;  
 format PoolArea best12.;  
 format PoolQC $2.;  
 format Fence $5.;  
 format MiscFeature $4.;  
 format MiscVal best12.;  
 format MoSold best12.;  
 format YrSold best12.;  
 format SaleType $3.;  
 format SaleCondition $7.;  
 input Id MSSubClass $  
 MSZoning $  
 LotFrontage LotArea Street $  
 Alley $  
 LotShape $  
 LandContour $  
 Utilities $  
 LotConfig $  
 LandSlope $  
 Neighborhood $  
 Condition1 $  
 Condition2 $  
 BldgType $  
 HouseStyle $  
 OverallQual OverallCond YearBuilt YearRemodAdd   
 RoofStyle $  
 RoofMatl $  
 Exterior1st $  
 Exterior2nd $  
 MasVnrType $  
 MasVnrArea ExterQual $  
 ExterCond $  
 Foundation $  
 BsmtQual $  
 BsmtCond $  
 BsmtExposure $  
 BsmtFinType1 $  
 BsmtFinSF1 BsmtFinType2 $  
 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating $  
 HeatingQC $  
 CentralAir $  
 Electrical $  
 "1stFlrSF"N "2ndFlrSF"N LowQualFinSF GrLivArea   
 BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr   
 KitchenQual $  
 TotRmsAbvGrd Functional $  
 Fireplaces FireplaceQu $  
 GarageType $  
 GarageYrBlt GarageFinish $  
 GarageCars GarageArea GarageQual $  
 GarageCond $  
 PavedDrive $  
 WoodDeckSF OpenPorchSF EnclosedPorch "3SsnPorch"N   
 ScreenPorch PoolArea PoolQC $  
 Fence $  
 MiscFeature $  
 MiscVal MoSold YrSold SaleType $  
 SaleCondition $;  
  
 if \_ERROR\_ then  
 call symputx('\_EFIERR\_', 1);  
  
 /\* set ERROR detection macro variable \*/  
run;  
  
/\* Stack the test and train datasets so the GLMSELECTS can learn the coefficients \*/  
/\* with the train data and predict the value of the log of the sale price with the test data \*/  
proc append data=test base=train force;  
run;  
  
proc contents data=train;  
run;  
  
proc print data=train (obs=10);  
run;  
  
data train\_6;  
 set train;  
 SalePrice\_l=log(SalePrice);  
 GrLIvArea\_l=log(GrLIvArea);  
run;  
  
ods graphics on;  
title "Stepwise Multiple Linear Regression - All vars No Ints";  
  
proc glmselect data=train\_6 plots(stepAxis=number)=(criterionPanel ASEPlot)   
 outdesign (addinputvars)=step\_allvars\_sel;  
 class MSZoning MSSubClass Street Alley LotShape LandContour Utilities   
 LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle   
 RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond   
 Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating   
 HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType   
 GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature   
 SaleType SaleCondition / param=ref;  
 model SalePrice\_l=MSZoning MSSubClass Street Alley LotShape LandContour   
 Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType   
 HouseStyle YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd   
 MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure   
 BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual   
 Functional FireplaceQu GarageType GarageYrBlt GarageFinish GarageQual   
 GarageCond PavedDrive PoolQC Fence MiscFeature YrSold SaleType SaleCondition   
 LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF   
 "1stFlrSF"n "2ndFlrSF"n LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath   
 FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces   
 GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch "3SsnPorch"n   
 ScreenPorch PoolArea MiscVal MoSold OverallQual OverallCond /   
 selection=stepwise(choose=cv select=sl) stats=press cvMethod=split(5)   
 cvDetails=all hierarchy=single;  
 output out=outData\_step\_allvars predicted=pred\_step\_allvars\_l;  
run;  
  
proc contents data=step\_allvars\_sel;  
run;  
  
proc reg data=step\_allvars\_sel plots=all;  
 model SalePrice\_l=&\_GLSMOD.;  
 ID id;  
 run;  
 title;  
 ods graphics off;  
 ods graphics on;  
 title "Forward Multiple Linear Regression - All Vars No Ints";  
  
proc glmselect data=train\_6 plots(stepAxis=number)=(criterionPanel ASEPlot)   
 outdesign (addinputvars)=forw\_allvars\_sel;  
 class MSZoning MSSubClass Street Alley LotShape LandContour Utilities   
 LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle   
 RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond   
 Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating   
 HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType   
 GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature   
 SaleType SaleCondition / param=ref;  
 model SalePrice\_l=MSZoning MSSubClass Street Alley LotShape LandContour   
 Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType   
 HouseStyle YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd   
 MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure   
 BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual   
 Functional FireplaceQu GarageType GarageYrBlt GarageFinish GarageQual   
 GarageCond PavedDrive PoolQC Fence MiscFeature YrSold SaleType SaleCondition   
 LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF   
 "1stFlrSF"n "2ndFlrSF"n LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath   
 FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces   
 GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch "3SsnPorch"n   
 ScreenPorch PoolArea MiscVal MoSold OverallQual OverallCond /   
 selection=forward(choose=cv select=sl) stats=press cvMethod=split(5)   
 cvDetails=all hierarchy=single;  
 output out=outData\_forw\_allvars predicted=pred\_forw\_allvars\_l;  
run;  
  
proc reg data=forw\_allvars\_sel plots=all;  
 model SalePrice\_l=&\_GLSMOD.;  
 ID id;  
 run;  
 title;  
 ods graphics off;  
 ods graphics on;  
 title "Backward Selection Multiple Linear Regression - All Vars No Ints";  
  
proc glmselect data=train\_6 plots(stepAxis=number)=(criterionPanel ASEPlot)   
 outdesign (addinputvars)=back\_allvars\_sel;  
 class MSZoning MSSubClass Street Alley LotShape LandContour Utilities   
 LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle   
 RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond   
 Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating   
 HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType   
 GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature   
 SaleType SaleCondition / param=ref;  
 model SalePrice\_l=MSZoning MSSubClass Street Alley LotShape LandContour   
 Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType   
 HouseStyle YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd   
 MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure   
 BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual   
 Functional FireplaceQu GarageType GarageYrBlt GarageFinish GarageQual   
 GarageCond PavedDrive PoolQC Fence MiscFeature YrSold SaleType SaleCondition   
 LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF   
 "1stFlrSF"n "2ndFlrSF"n LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath   
 FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces   
 GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch "3SsnPorch"n   
 ScreenPorch PoolArea MiscVal MoSold OverallQual OverallCond /   
 selection=backward (choose=cv select=sl) stats=press cvMethod=split(5)   
 cvDetails=all hierarchy=single;  
 output out=outData\_back\_allvars predicted=pred\_back\_allvars\_l;  
run;  
  
proc reg data=back\_allvars\_sel plots=all;  
 model SalePrice\_l=&\_GLSMOD.;  
 ID id;  
 run;  
 title;  
 ods graphics off;  
 ods graphics on;  
 title "Stepwise Selection Multiple Linear Regression - Most Promising Previously Selected Vars + Interactions";  
  
proc glmselect data=train\_6 plots(stepAxis=number)=(criterionPanel ASEPlot)   
 outdesign (addinputvars)=step\_topvarsints\_sel;  
 class MSSubClass Neighborhood RoofMatl Condition2 MSZoning SaleCondition   
 CentralAir KitchenQual/ param=ref;  
 model SalePrice\_l=MSSubClass|Neighborhood|YearBuilt|RoofMatl|BsmtFinSF1|GrLivArea|OverallQual|OverallCond|Condition2|MSZoning|GarageCars|TotalBsmtSF|SaleCondition|LotArea|CentralAir|KitchenQual|ScreenPorch  
 @2 / selection=stepwise (choose=cv select=sl) stats=press cvMethod=split(5)   
 cvDetails=all hierarchy=single;  
 output out=outData\_step\_topvarsints predicted=pred\_step\_topvarsints\_l;  
run;  
  
proc reg data=step\_topvarsints\_sel plots=all;  
 model SalePrice\_l=&\_GLSMOD.;  
 ID id;  
 run;  
 title;  
 ods graphics off;  
  
 /\* Macro to loop through creating each of the the three submission files \*/  
 %macro create\_submissions (method);  
 data &method.\_test\_stage1;  
 set outData\_&method.;  
  
 if Id ge 1461;  
 SalePrice=exp(pred\_&method.\_l);  
 run;  
  
 data &method.\_test\_stage2 (keep=Id SalePrice);  
 set &method.\_test\_stage1;  
 run;  
  
 /\* Routine to impute median value when predicted value is missing \*/  
 proc means data=&method.\_test\_stage2 median;  
 var SalePrice;  
 output out=tmp\_median\_SalePrice median=tmp\_median\_SalePrice;  
 run;  
  
 data \_null\_;  
 set tmp\_median\_SalePrice;  
 call symputx('tmp\_median\_SalePrice', tmp\_median\_SalePrice);  
 run;  
  
 %put &tmp\_median\_SalePrice.;  
  
 data &method.\_submission;  
 set &method.\_test\_stage2;  
  
 if SalePrice=. then  
 do;  
 SalePrice=input(&tmp\_median\_SalePrice., 11.2);  
 end;  
 run;  
  
 proc export data=&method.\_submission   
 outfile="/home/sarellano0/dataSets/&method.\_submission.csv"

dbms=csv replace;  
 run;  
  
%mend create\_submissions;  
  
%create\_submissions(method=step\_allvars);  
%create\_submissions(method=forw\_allvars);  
%create\_submissions(method=back\_allvars);  
%create\_submissions(method=step\_topvarsints);