Nivedha Thedsanamoorthy

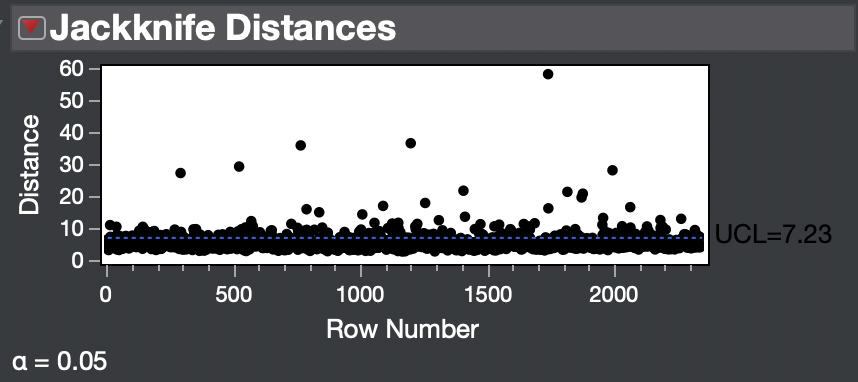
Hannah Carrico

Ryan Howe

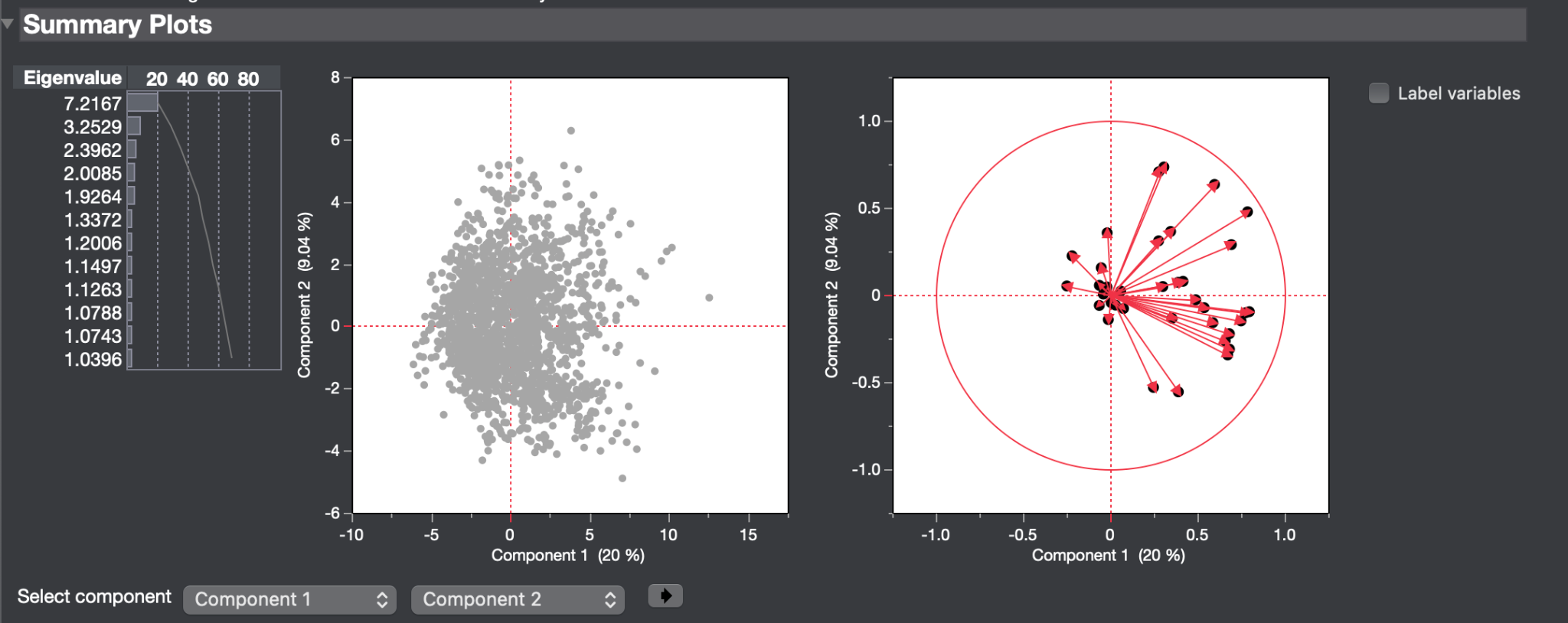
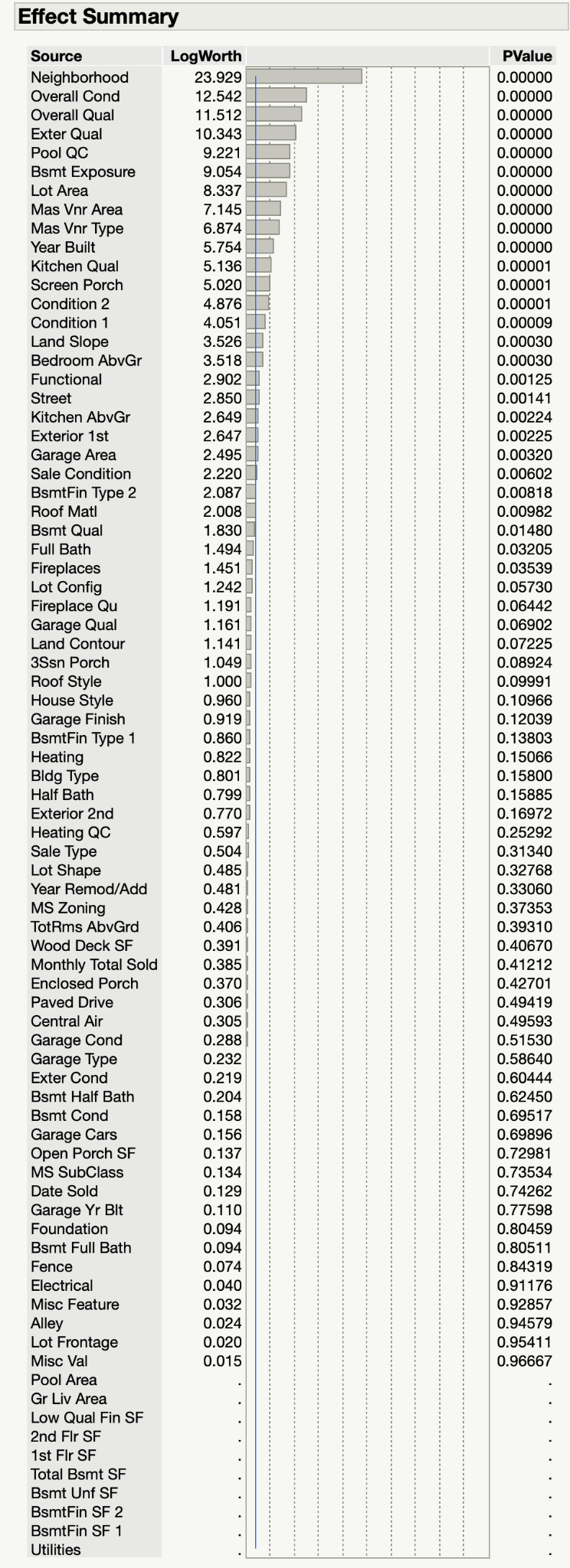
Anushreeya Gurung

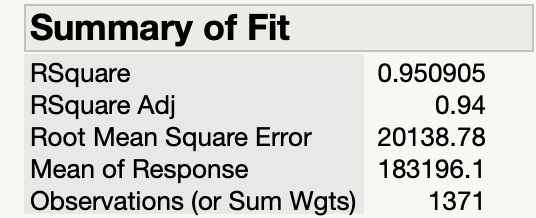
**Final Write-Up**

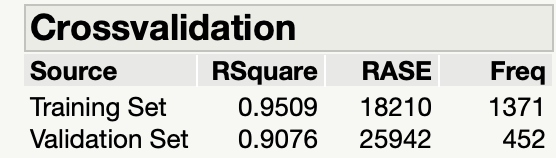
* **Wrangle Data** 
  + First an outlier analysis was conducted using the continuous variables. The outlier analysis used was performed through the multivariate platform.

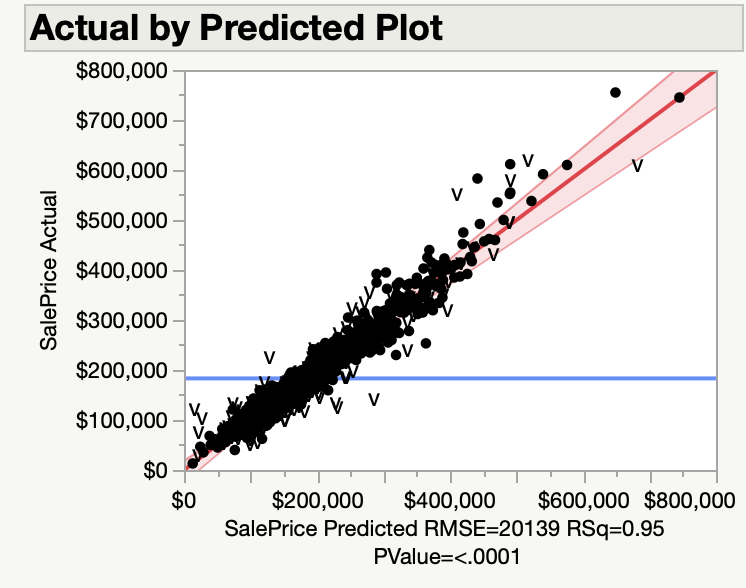


The Jackknife Distances plot shows outliers at row 291, 522, 765, 1199, 1741, and 1995. These 6 rows were removed for analysis. The Mahalanobis Distance (*Image 1*) and (*Image 2*) plots also showed these points as outliers.

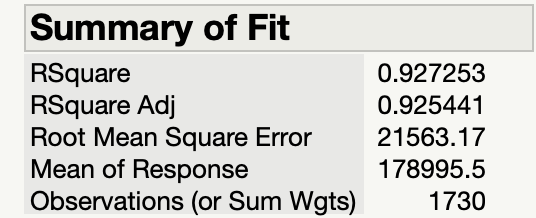
* **Explore Data** 
  + An analysis was done on missing values. It appears that missing data is not a problem in this data set. The Lot Frontage variable was missing 377 points but compared to the total number of observations (2344) it does not appear to be a problem. Many variables were missing no data. (*Image 3)*
  + A PC analysis was done on the data. Looking at the eigenvector for PC1, the highest predictor value is 0.293 for Gr Liv Area. Other significant predictors in PC1 are the three garage predictors, full bath, total bsmt SF and 1st Flr SF. This appears to represent predictors. None of these have values above 0.3. In PC2, the 2nd Flr SF has the biggest value at 0.407, and BedRoom AbvGr, and Total Room AbvGr also have values above 0.3. This PC seems to represent predictors about rooms above ground. 
* **Build Predictive Models and Validate the Models**
  + First, a standard least squares model was built. The RSquare value was .9407 which indicates good fit. The Actual by Predicted plot shows an outlier at row 1742 so this row was removed. A standard least squares model was then run again. This time the RSquare value was 0.9509. The validation RSquare value was 0.9076. This means the new model, with row 1742 removed, fits better. The effect summary is included to the side. The most important variables for sales price were Neighborhood, Overall Condition, Overall Quality, Exterior Quality, Pool QC, Bsmt exposure, Lot Area, Mas VNR Area, Mas VNR Type, and Year Built. These variables all had p-values of 0.0000 and high LogWorth values. Some variables were zeroed from the model. These variables were Pool Area, Gr Liv Area, Low Qual Fin SF, 2nd Flr SF, 1st Flr SF, Total bsmt SF, Bsmt Fin SF 2, BSMT Fin SF1, and Utilities. The actual by predicted plot also shows minimal lack of fit.

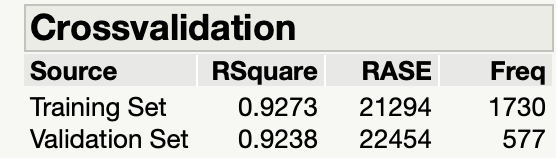




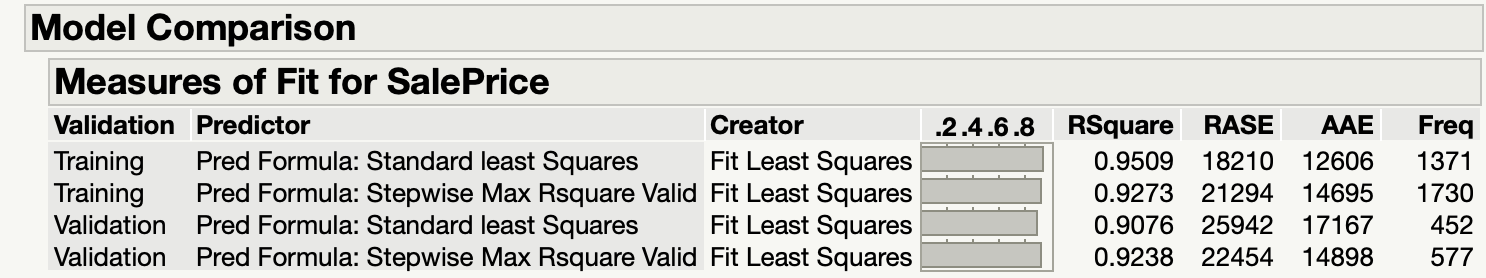


* + A stepwise model using Max Validation RSquare was fit. The RSquare was 0.9272 which is lower than the RSquare value for the standard least squares model. The validation RSquare value raised to 0.9238.

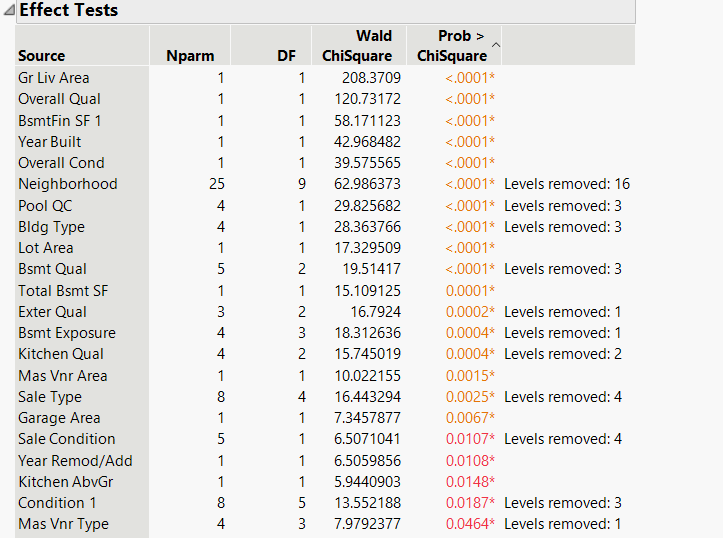




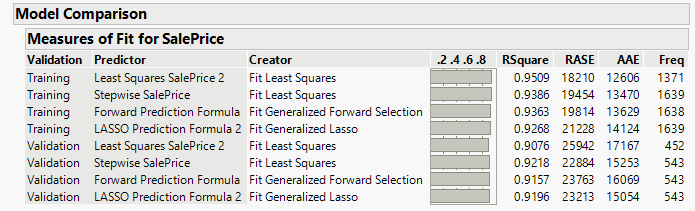
A model comparison was done between the standard least squares model and the Max Validation RSquare stepwise model.



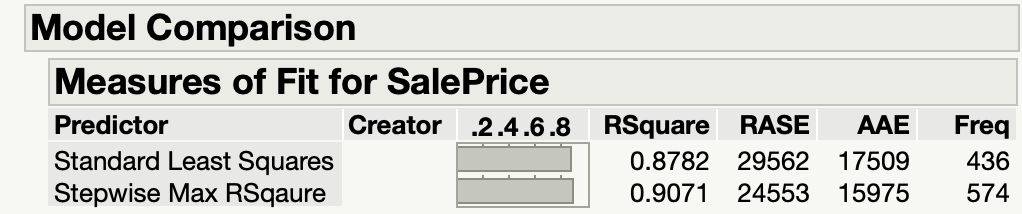
Based on the model comparison, the RSquare for validation was highest in the stepwise for Max Rsquare.

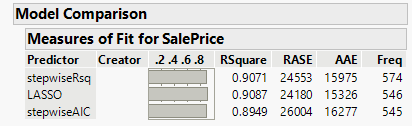
A Lasso model was run and the following is the output… 

For the lasso model, the most important predictors were GR Liv Area, Overall Qual, BsmtFin SF1, Year Built, Overall Cond, Neighborhood, Pool QC, Bldg Type, Lot Area, and Bsmt Qual. The actual by predicted plot shows minimal lack of fit and the RSquare value was 0.9357.

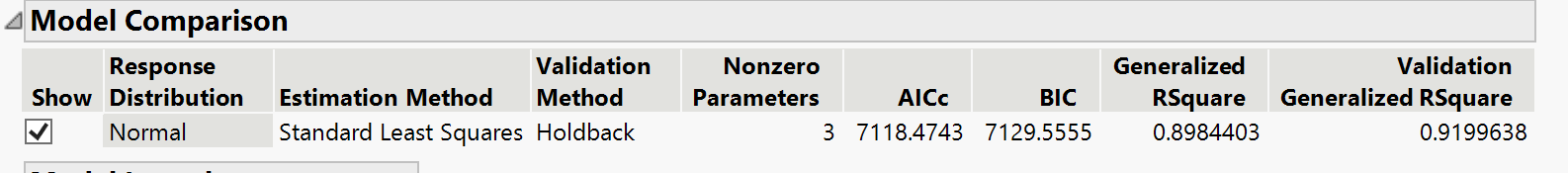
More models were run and model comparison was used. The models run were stepwise minimum AIC, generalized forward selection, and lasso. Based on the model comparison, the RSquare for validation was highest in the stepwise for Max Rsquare. At this point in time the stepwise for highest RSquare fit the data best.

* + Any analysis using the PC’s resulted in a R-square of less than .9 so PC’s were not used in this model image 4.
* **Deploy Models to New Data**
  + The prediction formulas were then used in the scoring data set. Model comparisons were then run.





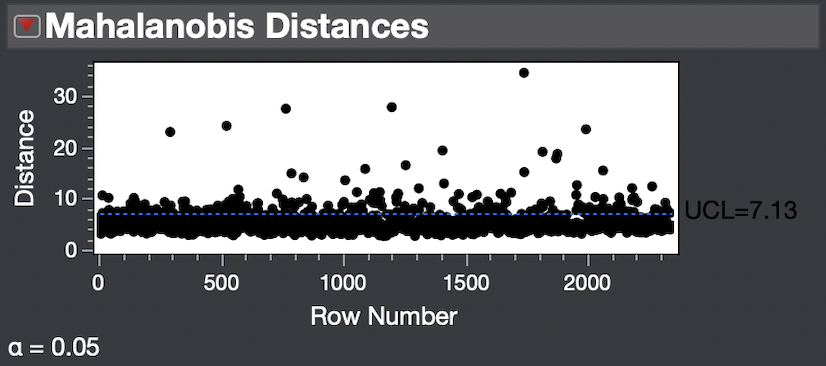




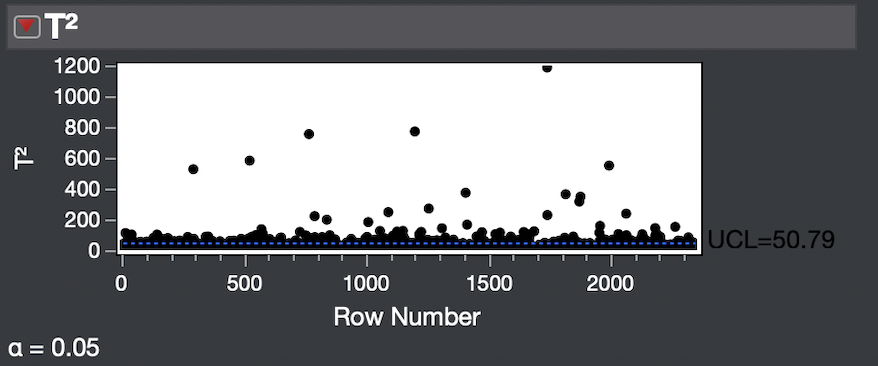
* **Conclusion**
  + In conclusion we determined the best model to be for the training data was the standard least squares model with an RSquare of 0.9509. The best model for the validation data was the stepwise max RSquare model with an RSquare of 0.9238. When applying all models to the scoring data the lasso model fits best with a validation RSquare of 0.9087. The stepwise max RSquare value fits second best with a RSquare value of 0.9071 which is close to the lasso model.
  + The most important predictors chosen by the lasso model were GR Liv Area, Overall Qual, BsmtFin SF1, Year Built, Overall Cond, Neighborhood, Pool QC, Bldg Type, Lot Area, and Bsmt Qual. These predictors were most useful in predictions with the scoring data.
  + Many predictors were determined to be useless and were removed by lasso, backwards and forwards selection. Notable among them, chosen by ANOVA for categorical variables agrees that Utilities will accept the null hypothesis that all the predictors have equal variance. This means that Utilities were not useful in predicting value house prices.

**Appendix**

*Image 1*



*Image 2*

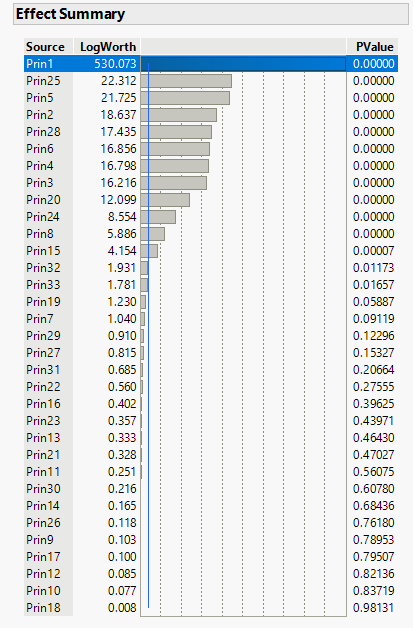
****

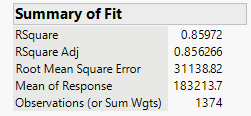
*Image 3*

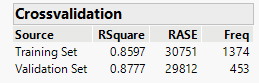




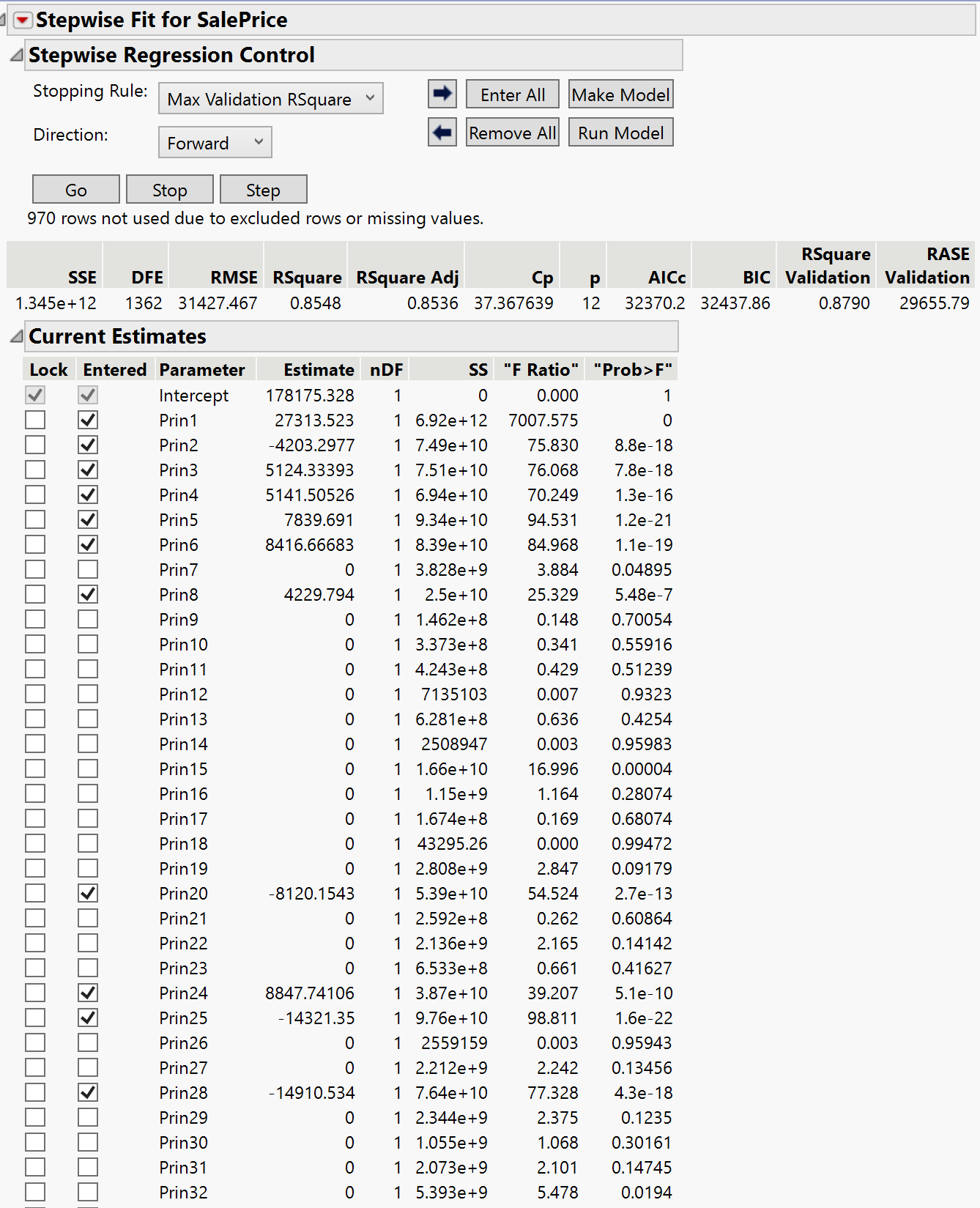
*Images 4,5,and 6 (r/t Standard Least Squares on PCs)*

****

****

****

*Image 7 (r/t Max Validation R-Squared on PCs)*

****

*Image 8 (r/t Anova)*

