**Introduction**

Our group has been hired on to examine 79 explanatory variables describing every aspect of residential homes in Ames, Iowa. Our client would first like us to examine 3 neighborhoods of interest (Northwest Ames, Brookside, and Edwards) and compare ground living area and its impacts on each house’s sales price in each of its respected neighborhood. Once providing our client with the information on sales price in the 3 desired neighborhoods, they would like us to build a predictive model for sales price for homes in all of Ames, Iowa. In order to successfully provided the most accurate data/model, our group will extensively examine all 79 explanatory variables and use a cross validation technique to provide our clients with important information.

**Data Description**

Our data was retrieved from the Ames Housing data set on Kaggle for a competition. The dataset contains 2930 observations divided into a testing and training set. The data set(s) consist of 79 explanatory variables that describe several aspects of a house which could have an influence on its sales price. For this dataset we cannot find more observations, however for general housing market data we can use several resources throughout different real-estate companies. Each variable used in the analysis was examined thoroughly with statistical evidence of collinearity. Our group used VIF scores, plots, excel spread sheets, p-values, and several other factors to decided which explanatory variable to keep or removed. For more information please refer to our appendix.

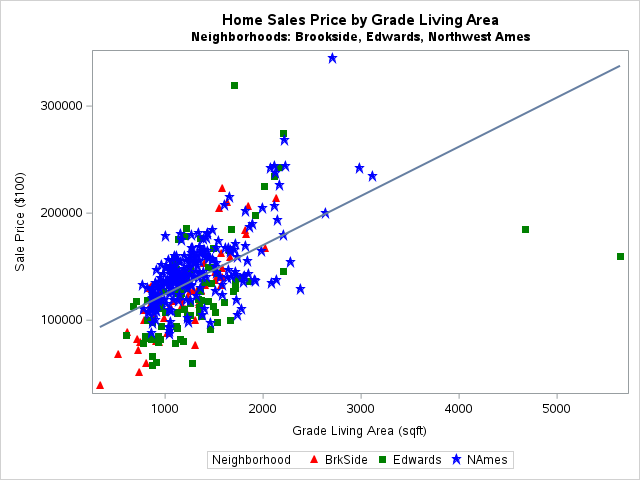
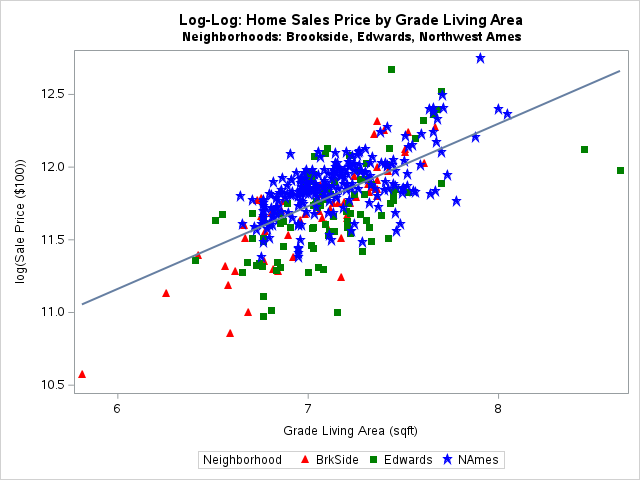
**Analysis Question 1**

**State the Problem:**

Century 21 Ames would like to obtain an estimate of how the sales price of a house was related to its square footage of the living area of the house; depending on its respective neighborhood of the Ames Iowa area.

**Plot and Review the Data:**

The scatter plot of the original data shows a limited linear relationship between the grade living area and sale price, with much of the data clustered between 1000 and 2000 sq ft. Both the Grade Living Area and Sale Price contain a couple influential datapoints possibly impacting the data (Appendix Figure 1). These datapoints were reviewed and determined that they were appropriate for the analysis. The neighborhood effect was associated with a 2B1 multiplicative increase in the median of sale price. A log-log transformation was used to control for the influential datapoints (Appendix Figure 4).



**Fit the Model:**

{log(SalePrice) | Neighborhood , log(GrLivArea)} = β0 + β1\*Neighborhood = Brookside + β2\*Neighborhood = Edwards + β3\*log(GrLivArea)

Pred {log(SalePrice)|Neighborhood, GrLivArea} = 7.9021 + 0.5558\*(Brookside) - 0.1328\*(Edwards) - 0.1532\* log(GrLivArea)

Pred {log(SalePrice)|Neighborhood=Brookside, GrLivArea} = 8.4579 - 0.1532\* log(GrLivArea)

Pred {log(SalePrice)|Neighborhood=Edwards, GrLivArea} = 7.7693 - 0.1532\* log(GrLivArea)

**Table

Description automatically generated** Table

Description automatically generated Table

Description automatically generated

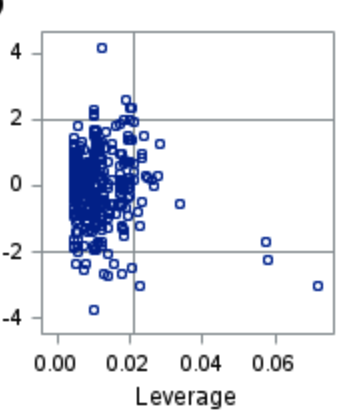
**Assumptions (log-log transformed data):**

* **Normality:** The q-q plot and the histogram show slight right skewness however not strong enough evidence against normality. There were a couple of influential observations with both high leverage and residual as well as high Cook’s D.
* **Linearity:** It was tough to check linearity in multiple dimensions however the scatterplot does show linearity with influential datapoints.
* **Constant Variance**: There was no visual evidence of differing standard deviation throughout the residual plot.
* **Independence**: Independence cannot be assumed so we will proceed with caution.

Chart, histogram

Description automatically generated Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Interpretation:**

There was overwhelming evidence to suggest the impact of a neighborhood has a 2B1 multiplicative increase impact on home sales price. Holding grade living area constant, it was estimated that the Brookside neighborhood sales prices was $87.57 (e-0.13279) more per 100 sq ft than the Northwest Ames neighborhood. A 95% confidence interval for this estimate was (e-0.18975, e-0.07583) = ($82.72, $ 92.70). Holding grade living area constant, it was estimated that the Edwards neighborhood sales prices was $85.79 (e-0.15323) more per 100 sq ft than the Northwest Ames neighborhood. A 95% confidence interval for this estimate was (e-0.19943, e-0.10703) = ($81.92, $ 89.85). Holding the neighborhood constant, it was estimated that the grand living area was $174.33 (e0.55579) per 100 sq ft with a 95% confidence interval for the estimate of (e0.49234, e0.61923) = ($163.61, $ 185.75).

**Conclusion:**

The real estate data for the Ames Iowa overwhelming showed that the sales price of a house was related to its square footage of the living area of the house and the neighborhood of the home (p-value < 0.0001). The Brookside neighborhood average sale price was the highest, followed by the Edwards neighborhood and then the Northwest Adams neighborhood.

**R-Shiny App**

Scatter plot of the sale price of neighborhood sale price vs the home square footage can be seen here: <https://nickmueller2.shinyapps.io/Neighborhood3/?_ga=2.68375953.127761547.1680987412-408874730.1680987412>

**Analysis Question 2**

**State the Problem:**

We would like to build the most predictive model for sale prices of homes in all of Ames Iowa.

**Exploratory Data Analysis:**

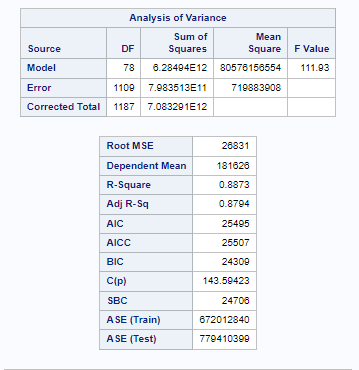
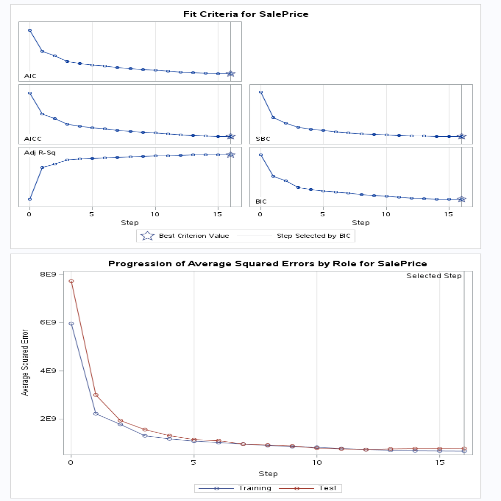
When exploring our data, we first started by examining the excel spreadsheet noticing obvious variables that could have a negative impact on our model. For example, we removed the variables SaleType, BsmtHalfBath, Functional, LowQualFinSF, BsmtFinSF2, ScreenPorch, EnclosedPorch, PoolArea, MiscVal, Utilities, and street since 85% to 99% of each column consisted of 1 main value. After examining the excel spreadsheet we inputted the data into SAS so that we could explore other competing factors against linearity. Factors considered included the p-value, VIF, sets with heavy outliers, unique identifiers (ex: ID), evidence against normality, and more. When addressing NA values, we decided that these were not “unknown” values, but values that were 0 (ex: The alley column NA= House didn’t have an alley way). Therefore, all NA/missing values we converted to 0 and other string values were converted to 1, 2, 3, and so on. As a team we also used the SAS Proc Corr to visually show numeric identifying variables which needed to be removed.

**Model Selection**

**Type of Selection: Stepwise**

For our stepwise model we were able to get a Kaggle score of 0.1637 which was also our best competing model in the Kaggle competition. Therefore, we used it for our custom model as well. The adjusted R square value was a .88 pr 88% with a cv press of 1.546512E12 and BIC score of 24309(figure 7 screenshots and code provided in appendix).



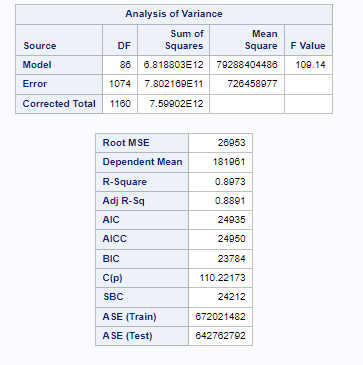
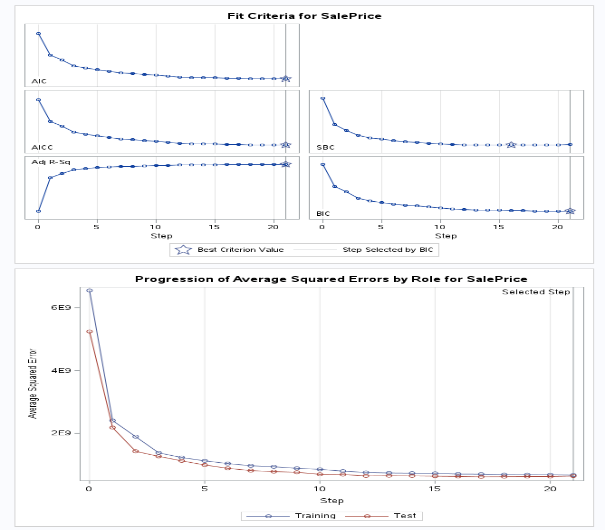


As we can see from the stepwise plot above our testing set seems to show a very close and similar path to the training set. There does not seem to be a large amount of variation, therefore we proceeded to submit out model to the completion for a Kaggle score.

**Type of Selection: Forward**

For our forward predicting model, we were able to get a Kaggle score of 0.16756. This Kaggle score was very close to our stepwise model, however, was not quick enough to pass it. The adjusted R square value was a .89 or 89% with a cv press of 1.607329E12 and BIC score of 23842 (figure 5 and 6 screenshots and code provided in appendix).



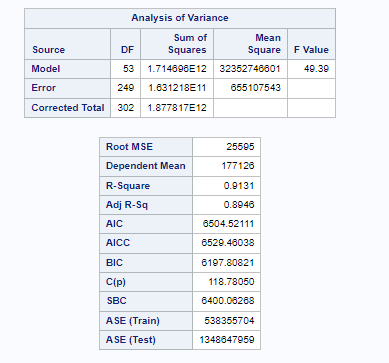
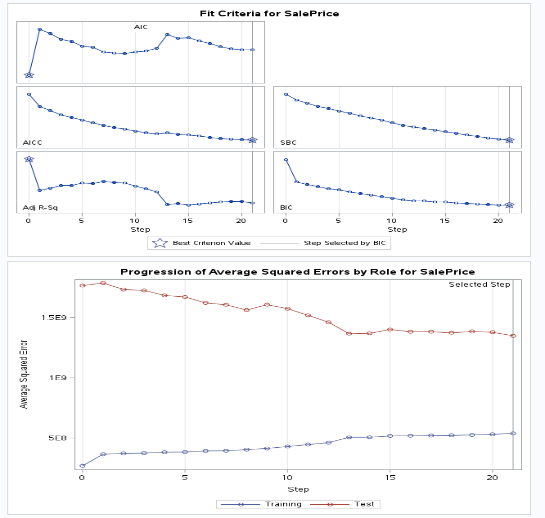


As we can see from the forward model plot above our testing set seems to show a very close and similar path to the training set. There does not seem to be a large amount of variation, therefore we will proceed with submission to the competition for a Kaggle score.

**Type of Selection: Backward**

For our backward model, it significantly underperformed compared to the other models. For unknown reasons the best Kaggle score we achieved with the backward model was a 0.38305. The adjusted R square values was .89 or 89% with a cv press of 2.650568E11 and a BIC score of 6197.8082123842 (figure 8, 9, and 10 screenshots and code provided in appendix).



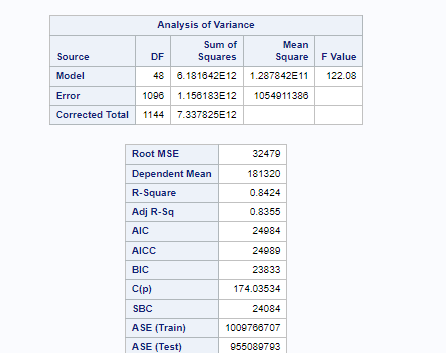
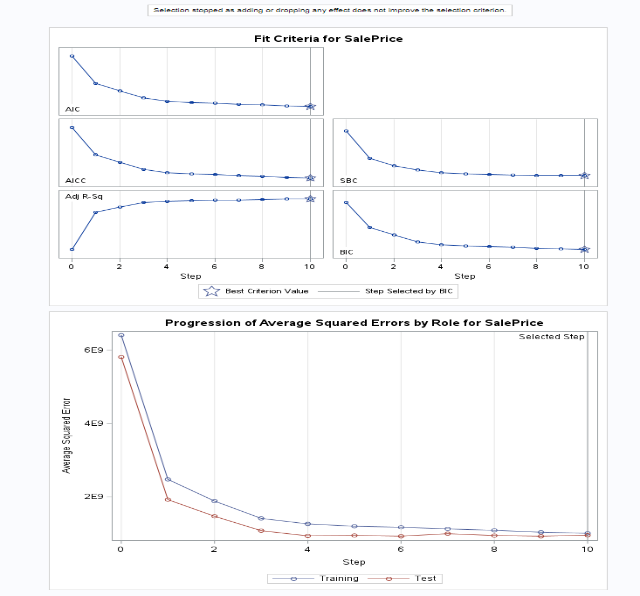


Unlike the forward and stepwise models, the backward model starts with every variable. As we can see from the plot above our model seems to trend together, however never intersects, showing us that the model is not quite perfect. That being said, with the time allotted this was our best seed attempt. Therefore, we will proceed with submission to the competition for a Kaggle score.

**Type of Selection: Custom**

For our custom model we used domain knowledge in the housing market to determine which variable would have the highest influence on sales price. Therefore, with the collected effort of cook’s D and other identifiers we were able to run a model that we personally thought would receive the best Kaggle score. However, our custom model only received a Kaggle score of 0.17581 which was worse than our original stepwise and forward model. The custom model also received an adjusted R square of 0.84 or 84% and a BIC of 23833(figure 11, 12 and 13 screenshots and code provided in appendix).





As we can see from the custom(stepwise) model plot above our testing set seems to show a very close and similar path to the training set. There does not seem to be a large amount of variation, therefore we will proceed with submission to the competition for a Kaggle score.

**Checking Assumptions**:

Once removing the values with visual evidence against normality we were able to obtain the residual plot, high leveraging point, and evidence against linearity. We were able to get plots that showed normality and collinearity.

* **Residual Plots:** We were able to see visual evidence of equal standard deviations with no evidence of curvature in the residual plot.
* **Influential point analysis (Cook’s D and Leverage**): We were able to remove high Cook’s D and leverage variable. With our final model Cook’s D we had the highest Cook’s D being around .2 to .3, however our team decided that they did not have enough leverage to be removed. We wanted to have as many variables as possible, so we proceeded with caution and took note of those higher Cook’s D 23842 (screenshots and code provided in appendix).
* **Address each Assumption:** Our group also reviewed a histogram that visually shows a normal bell shape curve. We also used a QQ plot to show linearity and possible outlier, for our final model we saw no evidence of significant outliers and majority variables seemed to follow a linear trend 23842 (screenshots and code provided in appendix).

**Comparing Competing Models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** | **BIC** |
| Forward | .89 | 1.607329E12 | 0.16756 | 23842 |
| Backward | .89 | 2.650568E11 | 0.3836 | 6197.80821 |
| Stepwise | .88 | 1.546512E12 | 0.1637 | 24309 |
| CUSTOM | .84 | 1.469603E12 | 0.17581 | 23833 |

**Conclusion:**

In conclusion, we observed 4 models which included a forward, backward, stepwise, and custom(stepwise) model. Our overall goal was to check the assumptions of the data, identify high leveraged values, and other factors that have a negative impact on our model. We were able to remove these high influential values. After removing the high influential values/variables we were able to run 4 successful models. In the end the best preforming model was the stepwise model with a Kaggle score of 0.1637 ranking our group number 2700.

**Analysis Question 1**

Figure 1: Proc SGSCATTER

Figure 2: Proc GLMSELECT Stepwise Model

A picture containing chart

Description automatically generated Graphical user interface, table

Description automatically generated

Figure 1 Figure 2

Figure 3: SGPLOT of Untransformed Data

Figure 4: SGPLOT of Transformed Data

Diagram, engineering drawing

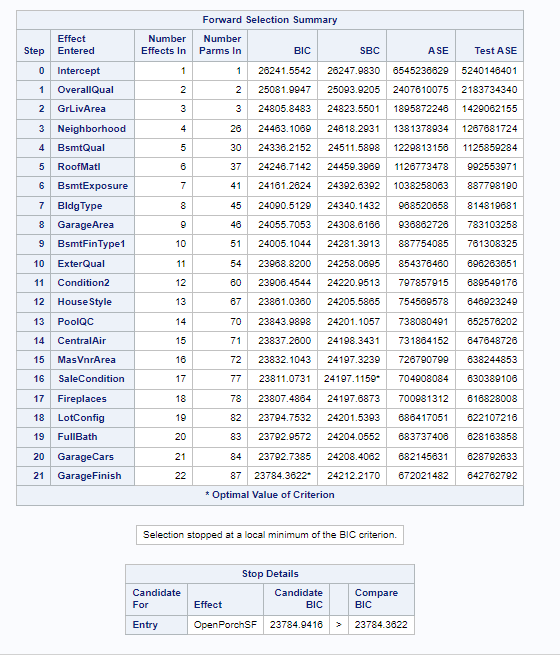
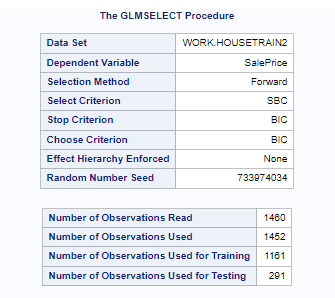
Description automatically generated Diagram, engineering drawing

Description automatically generated

Figure 3 Figure 4

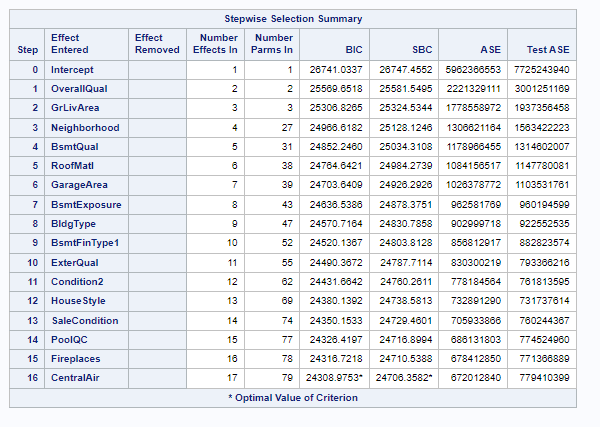
Figure 5: Forward Model Selection Summary

Figure 6: Forward Model Procedure

*Figure 5 Figure 6*

Figure 7: Stepwise Model Selection Summary

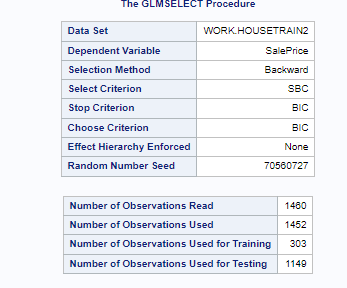
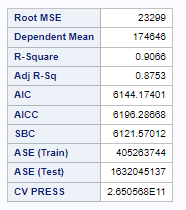


*Figure 7*

Figure 8: Backward Model Selection Summary

Figure 9: Backward Model Procedure

Figure 10: Backward Model performance statistics

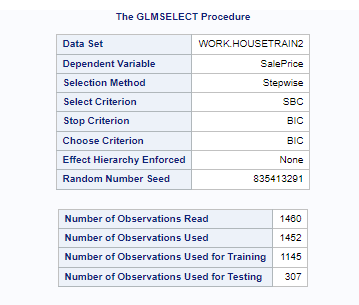
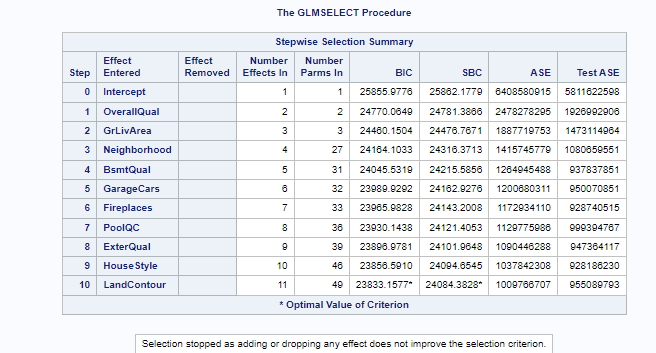
  

*Figure 8 Figure 9 Figure 10*

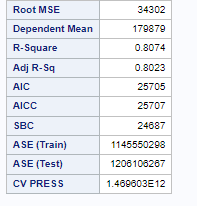
Figure 11: Custom Stepwise Model Procedure

Figure 12: Custom Stepwise Model Selection Summary

Figure 13: Custom stepwise Model performance Statistics

*Figure 11 Figure 12*



*Figure 13*