**Analysis 1** **Summary**: Century 21 Ames is seeking to analyze the relationship of the square foot living area and the sale prices of houses in the North Ames, Edwards, and Brookside neighborhoods in Ames, Iowa to improve their customer service and competitive edge.

**Problem**: Using multiple linear regression analysis estimate if there is a difference in the price compared to the living space square footage in the North Ames, Edwards, and Brookside neighborhoods and how those neighborhoods statistically compare with each other.

A graph of a number of black dots

Description automatically generated with medium confidenceA graph with a line

Description automatically generated**A graph of a graph showing a line and a red line

Description automatically generated with medium confidenceAssumptions Check**: An initial visual check of ***normality*** using histograms and QQ-plots show some right skewedness, probably resulting from outliers with strong leverage (See graphics below). Because the sample size is greater than 30 the Central Limit Theorem (CLT) is expected to resolve inconsistencies in normality. Nonetheless, we will log transform some of the data to check for improved fit. A visual review of the residuals also shows ***some variance across the line*** despite a good variance inflation factor (VIF) for sale price and living space (1.007636, 1.003811). The variables appear to be ***independent*** and have good ***linearity***. The primary issue with this test is the ***outliers***. Because the outliers have so much leverage, we opted to review these extremes to check for human error and other anomalies (see leverage and boxplots below).

**A graph with a number of houses

Description automatically generated with medium confidence**A graph of a graph with a line graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated**A graph with a number of boxes

Description automatically generated with medium confidenceInfluential Outliers**: A review of the outliers showed that IDs 524 and 1299 were new construction and in an incomplete status. We use Zillow to review the properties in Edwards associated with these IDs and concluded that the square footage exceeded any properties in that neighborhood and the low prices were probably related to lot price rather than a newly constructed home. Additionally, ID 534 in Brookside was assessed to be missing a zero on both sales (39300) and square footage (324 sqft). Lastly, ID 725 in Edwards had an extraordinary sale price (320000) for the living space (1698) and condition (5). All these outliers were determined to be the result of human error or reporting requirements for planned construction and were removed from the model.

**Comparing Competing models:** All models in analysis 1 used one generalized formula (salePrice ~ GrLivArea + Neighborhood) adjusted for the logged variable. We tested a multiple linear regression (r2 = 0.5081), log-linear (r2 = 0.4785), linear-log (r2 = 0.4897), log-log (r2 = 0.4903), and a logged interaction (r2 = 0.5094). The internal CV press for the log-log model returned a fairly good RMSE of 0.189923, suggesting strong performance, an r2 of 0.49082, suggesting decent predictive performance, and a MAE of 0.145467 indicating only minor difference between the predicted and actual values.

A graph of a graph

Description automatically generated with medium confidenceA graph of a number of black dots

Description automatically generated with medium confidenceA graph of a graph with numbers and lines

Description automatically generated with medium confidence**Log-Log transformations**: We opted to go with the Log-Log model with no interaction, primarily because it best met the assumptions. As you can see below, the log-log transformation and removed outliers improved ***normality***, ***linearity***, the ***residual distribution***, and ***variance*** across most neighborhoods, despite some right skewedness for living space in North Ames.

A graph of different numbers

Description automatically generated with medium confidence A group of graphs showing different types of numbers

Description automatically generated with medium confidence

A computer screen with white text

Description automatically generated**Parameters and Interpretation of the model**:

An analysis of the North Ames, Edwards, and Brookside neighborhoods shows relatively no statistically significant difference in the relationship of sales price and living area between Brookside (the reference neighborhood) and Edwards neighborhoods (p-value = 0.382). Although a purchase in the Edwards neighborhood is associated with a 0.02755 (or $97 non-logged) decrease in log-transformed price, this difference is not significant. We are 95 percent confident that the true logged sales price difference is found between (-0.0894956 and 0.0344004).

However, there is strong evidence (p-value < .0001) that the relationship between sales price and living area is different between Brookside and North Ames. For North Ames, there is evidence, holding all other variables constant, that purchasing in that neighborhood is associated with an estimated 0.12412 ($113 non-logged) increase in the log-transformed sale price for each unit of living space compared to Brookside. We are 95 percent confident that the true logged difference in the purchasing price between North Ames and Brookside is between (0.06889 and 0.179357).

A graph of a graph showing a number of living area

Description automatically generated with medium confidenceRegarding the relationship between sales price and living area across all values, there is overwhelming evidence (p-value <0.0001) of a positive linear relationship. For each log-transformed unit of living space, the log-transformed sales price is increased on average by 0.57229 units. In real world terms, that is for every 100 square feet of living space there is an estimated increase of $171 in the base purchase price.

**Conclusion**: There is overwhelming evidence (p-value <0.0001) of a positive linear relationship between the sales price and the square footage of living space. However, this relationship does not mean there are significant differences between North Ames, Edwards, and Brookside neighborhoods in purchasing price and living space. A breakdown of the price to square footage between Edwards and Brookside shows no significant difference. Our best estimate was that square footage cost $97 dollars less in Edwards than Brookside, but this was not statistically significant. In North Ames, however, there was strong evidence that square footage cost $113 more per hundred square feet than the same living space in Brookside. This model explained only 49.03 percent of the variation and other variables should be considered to strengthen the model. (See Appendix 1)

**Analysis 2** **Summary**: Expanding on the request from the customer, we will seek to build the most predictive model for the sale prices of homes in all of the neighborhoods in Ames Iowa.

**The Problem**: Build three models; a simple linear regression model, a multiple linear regression model comparing sales price, above ground living Area (GrLivArea), and number of full baths, and a custom multiple linear regression model that will show the relationship of the most predictive variables. These models will be compared to each other using metrics such as adjusted R squared, CV press and Kaggle score.

**A screenshot of a graph

Description automatically generatedA graph with a line and a line

Description automatically generated with medium confidenceModel 1: Simple Regression – Examining the relationship between Salesprice and lot area.**

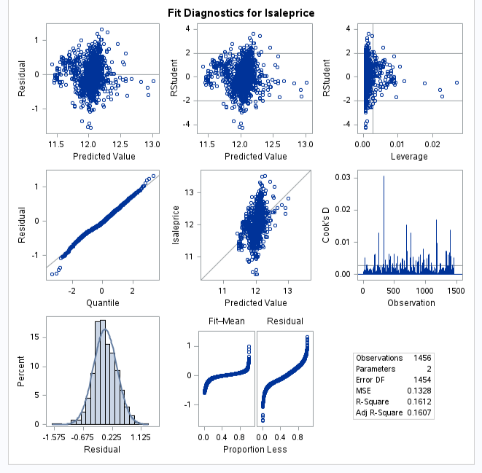
**Assumptions Check**

An initial visual review of the plots between saleprice and lot area shows some right skewedness in the histogram and q-q plot, suggesting the assumption for normal distribution is not met. Additionally, there appears to be clustering of the residuals and several outliers with heavy leverage, suggesting that there is not equal spread. In regard to the sample, we assume that the observations are independent and that the data represents the entire single family housing population in Ames.

**Log-Log Transformation**

**A graph of a scatter plot

Description automatically generated**A diagram with numbers and symbols

Description automatically generated with medium confidence****To mitigate for the lot size outliers (cases with 2 to 5 acres) we will perform a log-log transformation on the sales price and lot area and examine the relationship between the two logged variables. (Note: Limited information on the cost associated with acreage precluded an analysis that would have allowed us to drop these most egregious outliers. Instead of ranged exclusion, we opted for log-log transformations.) By first plotting the log - log transformed data, there is a better fitting linear relationship between the logged saleprice (response variable) and the explanatory variable (Lot area). Furthermore, a first look at the graphics seem to meet the assumptions. There does not appear to be any potential influential point. Judging from the scatter plot, q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution with constant variance. We continue to assume here that the observations are independent.

A screenshot of a computer

Description automatically generated

Pred log(Saleprice) = 𝛽0 + 𝛽1 log(LotArea)

Pred log(Saleprice) = 9.193 + 0.311log(LotArea)

We are 95% confident that for each doubling of the lot area, the median Saleprice rate will increase between approximately (2^0.27) 20.6% and (2^0.35) 27.5%. Our best estimate is an increase of (2^0.31) 24%.

**Model 2: Multiple Regression with SalePrice~GrLivArea + Full Bath**

By plotting the data first, we can tell there is some evidence of a positive relationship between the explanatory variables and the response variable. Please see appendix 2 for additional visual plots.

**Assumptions check:**

There does appear to be an influential point in the Rstudent plot. That observation will require further investigation. Judging from the scatter plot, q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution but there is not enough evidence to indicate a constant variance. We assume here that the observations are independent.

A close-up of a graph

Description automatically generated

**Log – Log Transformation**

We will do a log transformation of Saleprice and living area and plot those variables against each other to see if there is a linear relationship between them. Please see appendix 2 for plot.

Looking at the scatterplot matrix there is a positive linear relationship between the explanatory variables (Living area and full bath) and the response variable (Sales price). The log transformed variables show a more positive linear relationship between lsaleprice and lGrLivArea than the original data. We will build and fit a model with the log transformed data.

**Assumptions Check**

There do appear to be three influential points. That observation will require further investigation.

Judging from the scatter plot, q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution and a constant variance.

We assume here that the observations are independent.

We fit our model with ran on our explanatory (lsaleprice) and dependent (lgrlivArea and Full bath). This model has an Adjusted R square of 0.5637 which means this model estimates about 56.37% of the variation in salesprice is explained by the explanatory variables, a SBC of -3860.2586, and a CV press of 101.20 . Please see appendix for more details on the model.

The next step is to build a model with the log of sale price as the explanatory variable and fit the model.

***pred{𝑙saleprice}= 𝛽\_0+𝛽\_1 𝑙grlivArea +𝛽\_2FB0 +𝛽\_3 FB1 + 𝛽\_4 FB2 Reference = FB3***

***pred{𝑙saleprice}= 6.9797+0.723 𝑙grlivArea - 0.2424FB0 - 0.2992FB1 – 0.142 FB2***

***pred{𝑙Saleprice ┤|𝑙GrLivArea, FullBath=0}=6.9797+0.723 𝑙grlivArea - 0.2424FB0 = 6.9209 + 0.723 𝑙grlivArea***

***pred{𝑙Saleprice ┤|𝑙GrLivArea, FullBath=1}=6.9797+0.723 𝑙grlivArea - 0.2992FB1 = 6.6805 + 0.723 𝑙grlivArea***

***pred{𝑙Saleprice ┤|𝑙GrLivArea, FullBath=2}=6.9797+0.723 𝑙grlivArea – 0.142 FB2 = 6.8377+ 0.723lgrlivArea***

***pred{𝑙Saleprice ┤|𝑙GrLivArea, FullBath=3}=6.9797 + 0.723 𝑙grlivArea***

We are 95% confident that for each doubling of the living area, the median Saleprice rate will increase between approximately 59% (2^0.67 = 1.59) and 71% (2^0.78=1.71). Our best estimate is an increase of 65 % (2^0.72 = 1.647) after holding all other variables (Full Bath) constant. To get the regression equation for a specific full bath number, we will need to adjust that full bath’s coefficient with the A screenshot of a computer screen

Description automatically generatedintercept in our regression model.

**Custom Model**

A screenshot of a computer screen

Description automatically generatedFor the custom model, we decided to go with an automatic selection technique and then finetune the model with manual selection techniques. We did this by running stepwise, forward and backward selection models on the explanatory variables in the housing dataset and our response variable (Saleprice). The stepwise method made the most sense with the highest R score and lowest CV press. (See the graphic for Adjusted R square, SBC, and CV press). So, for our custom model we chose the stepwise technique.

Because our model was so large, we opted to reduce the number of variables manually by removing one variable at a time and retesting the model. We will fit our final model (lSalePrice ~lGrLivArea + OverallQual + MSSubClass + GarageCars + YearRemodelAdd + Neighborhood). Several of our variables were categorical and their coefficients require individual assessment, which we review in appendix 2.

**Assumptions Check**

Looking at our fit residuals, there does not appear to be any influential point on the residual plots based on our RStudent Plot. Judging from the scatter plot, q-q plot and there is no evidence that the residuals do not follow a normal distribution and a constant variance although our histogram looks slightly skewed to the left due to the presence of some outliers in the data.

We will assume that all observations are independent of each other.

(Please see appendix for fit residuals plot)

**Comparing Competing Models**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Simple Linear Regression | .16 | 103.5 | .983 |
| Multiple Linear Regression | .56 | 101.2 | .864 |
| Custom MLR Model | .899 | 31.53 | .552 |
| Other Models … | … | … | … |
|  |  |  |  |

**Conclusion**

Comparing the three models, that is our simple Linear regression, our multiple linear regression (MLR), and the Custom MLR model, it is quite obvious that the custom MLR model with lGrLivArea + OverallQual + MSSubClass + GarageCars + YearRemodelAdd + Neighborhood as predictors for salesprice is the most useful model to predict the sale price of the homes in Iowa. It has the highest adjusted R2 of 0.899, and the lowest CV press of 31.53, our chosen metric for comparison. Approximately 89.9% of the variation in the salesprice is explained by explanatory variables in the model. Interpreting the slope for our regression equation will depend on what slope we want to interpret. For instance, interpreting the slope for living area will mean that we are 95% confident that for each doubling of the living area, the median Saleprice will increase between approximately 38% (2^0.46 = 1.38) and 45% (2^0.54). Our best estimate is an increase of 41 % (2^0.5 = 1.41) after holding all other variables constant. There is strong evidence (p-value < .0001) of a relationship with sales price.

Additionally, several categorical variables showed a statistically significant relationship with sales price (See appendix 2), including several neighborhood, Model type, and the number of cars to the garage, but varies based on the category. To get the regression equation for a specific categorical variables, we will need to adjust that variable’s coefficient with the intercept in our regression model.

Project RShiny App: [Statistical Insights (weiprecht.github.io)](https://weiprecht.github.io/index.html)

Project RShiny Page: [SFDS Final Project (weiprecht.github.io)](https://weiprecht.github.io/Housing_Final_Project_Code.html)

Interactive Graphic: [Ames Housing App (shinyapps.io)](https://weiprecht.shinyapps.io/ames_housing_app/)

Project RShiny Github: [Ames Housing App (shinyapps.io)](https://weiprecht.shinyapps.io/ames_housing_app/)

**Appendix 1: Analysis 1 Code** –

A graph of different plots

Description automatically generated with medium confidenceA computer screen shot of a black screen

Description automatically generatedInitial analysis of original data

A computer screen shot of a program

Description automatically generated

A map of a neighborhood

Description automatically generatedZillow maps of Brookside, Edwards, and North Ames neighborhoods.

A map of a neighborhood

Description automatically generatedA map with blue lines

Description automatically generated

**Internal Cross Validation:**

**A screenshot of a computer

Description automatically generatedA computer screen shot of a program

Description automatically generated**

**A computer screen with white text

Description automatically generatedA screenshot of a computer program

Description automatically generatedInteraction code:**

**A screen shot of a computer program

Description automatically generatedScatterplot code:**

**A screenshot of a computer program

Description automatically generatedConversion code (transform coefficients back to original.)**

**A graph of a number of lines

Description automatically generated with medium confidenceA graph with numbers and letters

Description automatically generatedA graph of a sales graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generatedA screenshot of a computer program

Description automatically generatedAppendix 2: Initial dataset for Analysis Question 2**

**A graph of a number of people

Description automatically generatedAdditional Data for Model 1:** This is the code for the sales price by the lot areas. The code includes a cv press details and then the simple linear model.

proc glmselect data = housing3;  
/\* where LotArea not in (63887,40094); \*/  
model lSalePrice = llotArea / selection= Stepwise(stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq;  
run;

/\* Regression Code to get plots for SLR model \*/

proc reg data = housing3;

A screenshot of a computer

Description automatically generatedmodel lsaleprice = lLotarea;

run;

A screen shot of a graph

Description automatically generatedA blue dot diagram with white text

Description automatically generated

**Code for Model 2: Includes CV Press and output.**  
/\* Backward Selection for 2nd Model \*/ proc glmselect data = housing3; Class Fullbath; model lSalePrice = lGrLivArea FullBath / selection= backward(stop = cv) cvmethod = random(5) CVDETAILS stats = adjrsq; /\* selection = stepwise(stop = SL SLE = 0.05 SLS =0.05) STATS=adjrsq; \*/ run;

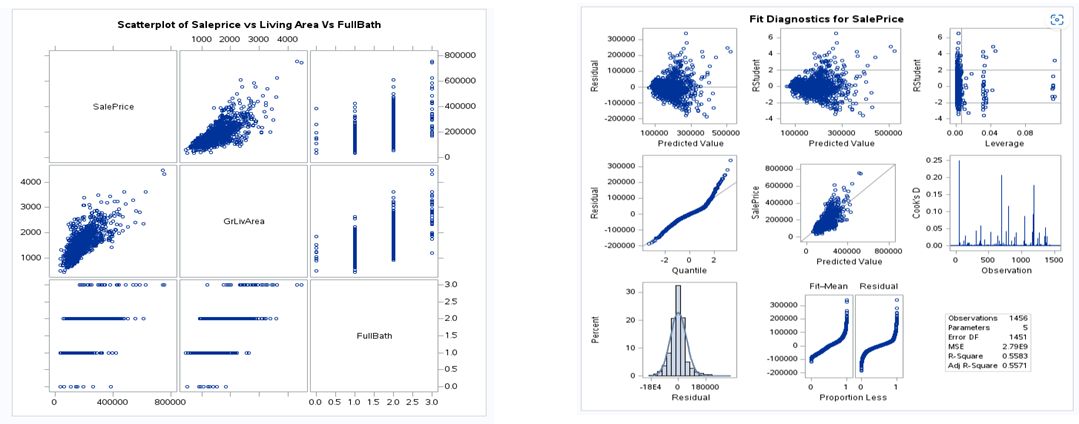
/\*Proc glm model to fit 2nd Model\*/

proc glm data = housing3 plots= all;

Class Fullbath;

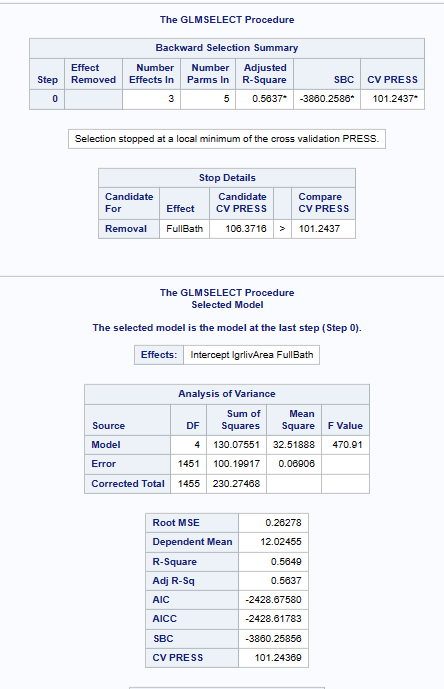
model lSalePrice = lGrLivArea Fullbath / solution clparm;

run;



A close-up of a graph

Description automatically generated



A diagram of a graph

Description automatically generated with medium confidenceA blue dot diagram with numbers

Description automatically generated with medium confidence

Parameter estimates for model 2:

A table with numbers and symbols

Description automatically generated

**Code for Model 3:**

**Stepwise Selection For custom model**

proc glmselect data = housing3;

Class MsSubclass Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual OverallCond 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;

model lSalePrice = MSSubClass LotFrontagen lLotArea 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 BsmtFinType2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical fstFlrSF ScndFlrSF LowQualFinSF lGrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch SsnPorch3 ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition / selection=stepwise (stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq;

/\* selection=backward (stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq; \*/ /\* selection = backward(stop = SL SLS = .01) stats = adjrsq; \*/

run;

**Building and fitting custom model with proc glm**

proc glm data=housing3 alpha=0.05 plots=all; class Neighborhood OverallQual MsSubclass GarageCars ;

model lSalePrice = lGrLivArea OverallQual MsSubclass GarageCars YearRemodAdd Neighborhood / solution clparm; run;

**Proc Glmselect on custom model to get the CV press score**

proc glmselect data=housing3;

class Neighborhood OverallQual MsSubclass GarageCars ;

model lSalePrice = lGrLivArea OverallQual MsSubclass GarageCars YearRemodAdd Neighborhood / selection=stepwise (stop = cv) cvmethod = random(10) CVDETAILS stats = adjrsq;

run;

Plots and Parameter Estimates for the custom model 3:

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

Final Custom Model Coefficients:

