SOUTHERN METHODIST UNIVERSITY
MSDS 6371(401)

Kaggle Project

Michael J Wolfe
Sandesh Ojha
Carl Walenciak
Travis Daun

Github Repository

https://github.com/mjwolfe91/SFDS_401_Team3_Kaggle_Project

February 23, 2019

Contents

1. Introduction

Many factors can impact the sale price of residential real estate. This report will explore those various aspects to help define what factors tend to impact home prices. We start by looking at the three distinct neighborhoods (North Ames, Edwards, and Brookside) in Ames, Iowa that Century 21 Ames sells houses to help better understand how above ground living space relates to sales price in each of these neighborhoods and if the neighborhood has an impact on the sales price. At the conclusion of this analysis, we provide a conclusion that quantifies the relationship between living area and sales price with respect to these neighborhoods.

After examining the impact of above ground living space on sales price for these three neighborhoods, we provide a much more complex analysis of factors that can be used to determine sales price for Ames as a whole. While we will use various methods to determine what factors are important for our predictive model, we will test this model on a blind dataset to show how it performs. The results of this model on the test data will be provided to further build the level of confidence in this tool for predicting future home sale prices.

2. Data Description

The dataset used for this analysis was retrieved from www.kaggle.com/c/house-prices-advanced-regression-techniques.

DATA The data for this evaluation contained 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. A complete description of these data can be found in Appendix D. The explanatory variables contain both categorical and numeric attributes. Appendix C.1, provides a high level summary of the variables and variable types contained in this dataset.

PREPROCESSING The variable `LotFrontage` posed a challenge since it was a continuous numerical variable that contained `NA`. This is because the variable was for the linear feet of street connected to property. In many cases (259 of 1460) this was either unknown or unrecorded. Our team made the decision to convert these NA values to 0. We believe this is an acceptable practice since we performed a sensitivity analysis by replacing the NAs with values from 0 to 140 (mean value is 70) with no impact on the linear regression model selection process and this factor was not utilized in any of our final models.

TRAINING Training of the linear regression models will be done utilizing the `training.csv` data obtained from above. A five fold cross validation will be employed for model selection.

TESTING Testing of the linear regression models will be done utilizing the `test.csv` data obtained from above.

RESULTS Datafiles containing the test results of the linear regression models can be found in our Github repository at https://github.com/mjwolfe91/SFDS_401_Team3_Kaggle_Project

3. Analysis Question 1

3.1. Restatement of Problem

Request Century 21 Ames only sells houses in the Noth Ames, Edwards and Brookside neighborhoods and wants an estimate of how the sale price of the house is related to the square footage of the living area of the house. Additionally, Century 21 Ames would like to know if the sale price (and its relationship to square footage) depends on which neighborhood the house is located in. A fit a model will be used to answer this question.

Deliverable Provide the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. Provide evidence that the model assumptions are met and that any suspicious observations (outliers / influential observations) have been identified and addressed. Finally, a conclusion that quantifies the relationship between living area and sale price with respect to these three neighborhoods.

3.2. Build and Fit the Model

Appendix A. Source Code For Analysis 1 contains the SAS code used to check the assumptions, clean the data, and run the model to determine the best estimate for sale price based on square footage for the North Ames, Edwards, and Brookside neighborhoods. We removed four data points from the original dataset due to being outliers. First sale prices greater than $300,000 were determined to not be representative of the total population of these three neighborhoods. Second, sale condition was limited to only those that were normal sales. Again we feel that home sales that were not normal sales such as foreclosures, linked properties, and land purchases were not representative of the population of interest.

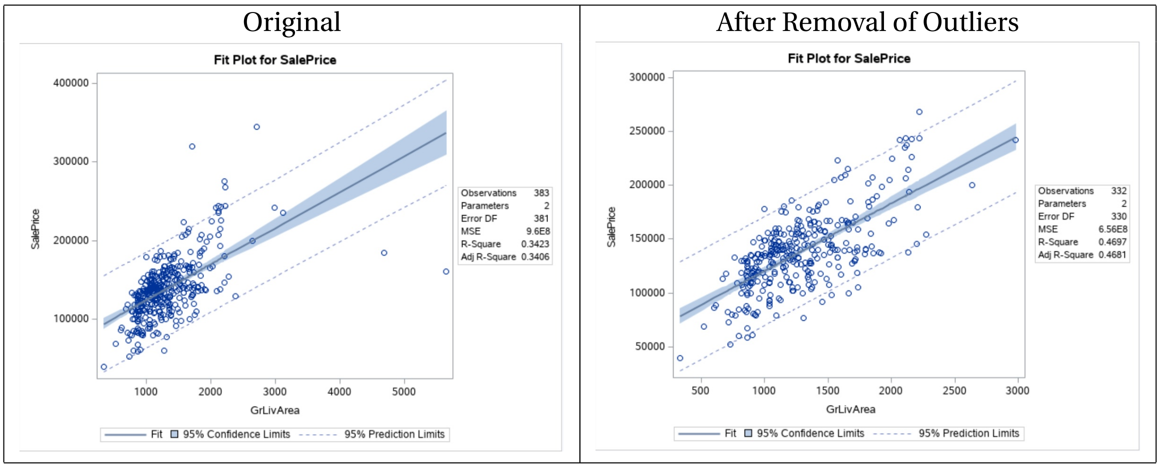

Figure 3.1: Simple Linear Regression Models.

3.3. Checking Assumptions

3.3.1. Assumptions

Linearity The linearity assumption is met by the reviewing the scatter plots associated with data. Fig 3.1 shows a plot of SalePrice vs. GrLivArea and by removing the outliers the linearity assumption is reasonably met.

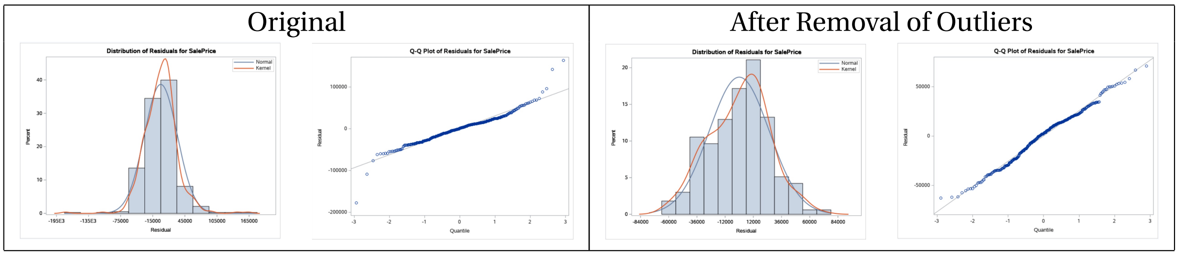

Figure 3.2: Normality Plots.

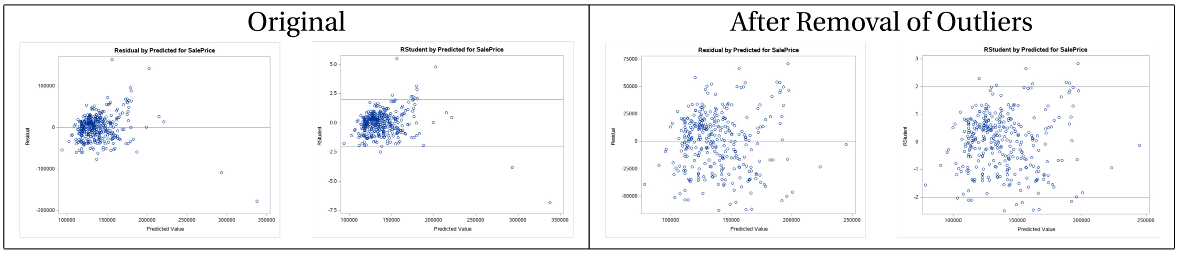

Figure 3.3: Residual Plots.

CONSTANT VARIANCE The residual plot, Fig 3.3 resembles somewhat of a random scatter of points around the 0 line, although there is a slight suspicion of non-constant variance judging from the dense cloud around the predicted value of $130,000. Also shown is the Studentized Residual Plot which is very similar to the residual plot, although this plot identifies potential outlying observations.

NORMALITY Based upon the histograms and q-q plots in Fig 3.2 there is no evidence to suggest that normality of the data. Additionally the random scatter associated with the residual plots in Fig 3.3 also support the normality assumption.

INDEPENDENCE The independence assumption can be assumed to be maintained since these are all unique sales in a free housing market.

3.3.2. Outliers and Influential Points

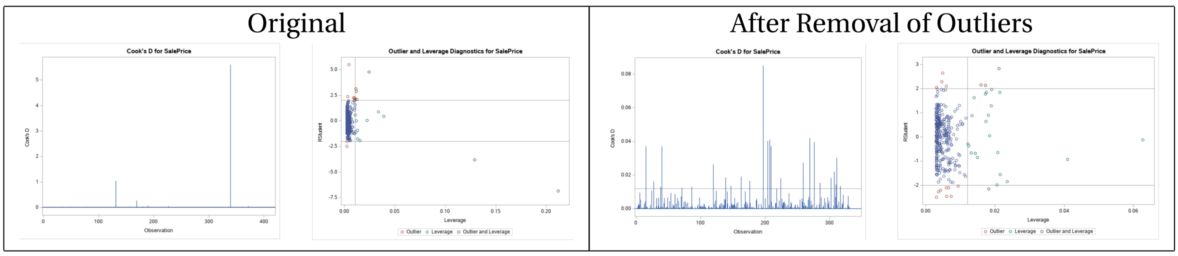

Figure 3.4: Influential Point Plots.

OUTLIERS / INFLUENTIAL OBSERVATIONS There are distinct outliers that can be seen both in the Cook's D and Outlier and Leverage Diagnostics seen in Fig 3.4. By removing the observations that resulted from non-normal sales conditions such as foreclosures and sale prices that were significantly outside the population, the Cook's D and Outlier and Leverage Diagnostics significantly improved. We are confident that the removal of these observations was appropriate since they do not represent the population as a whole. Additionally we are confident that the remaining data does not contain significant outliers or leverage points that need to be addressed further.

The model is a reasonable fit without transformations. The removal of observations not reflective of the population, such as non-normal sale conditions and home sales greater than \$300,000, seems appropriate and enables the required model assumptions to be met.

3.3.3. Effect by Neighborhood

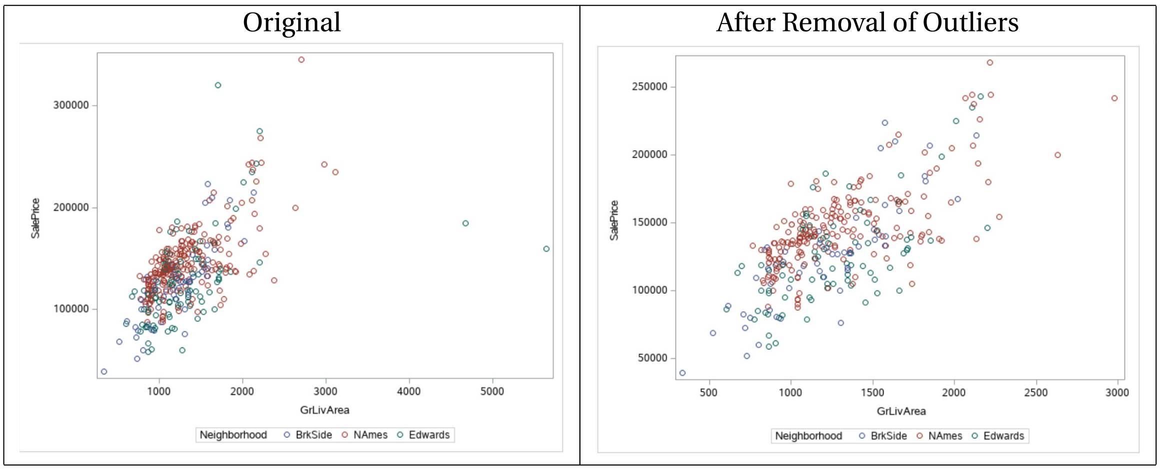

Figure 3.5: Scatterplot of Sale Prices vs Living Area by Neighborhood.

MODEL A model was developed to individually look at the neighborhoods of interest to determine if the sale price is impacted by neighborhood. Fig 3.5 shows the sales price vs living area by neighborhood. The model shows (Table 3.1) that the intercept and slope for Brookside and Edwards do differ from North Ames with statistical significance (p-values: < .0001 and .0077). As such we have determined to choose this model which allows for different intercept and slopes based upon the neighborhood. The resultant model can be seen in Fig 3.6.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74982.03784	B	5925.64914	12.65	<.0001	63324.70074	86639.37493
GrLivArea100	54.93056	B	4.36399	12.59	<.0001	46.34542	63.51570
Neighborhood BrkSide	-55776.16653	B	12012.18226	-4.64	<.0001	-79407.34256	-32144.99051
Neighborhood Edwards	-30273.99848	B	11296.53114	-2.68	0.0077	-52497.29731	-8050.69965
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	33.31303	B	9.33444	3.57	0.0004	14.94970	51.67637
GrLivArea*Neighborho Edwards	8.24334	B	8.45713	0.97	0.3304	-8.39409	24.88077
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

Table 3.1: Results of Neighborhood Impact on Sales Price

3.4. Model Metrics

Model metrics can be seen in Table 3.2.

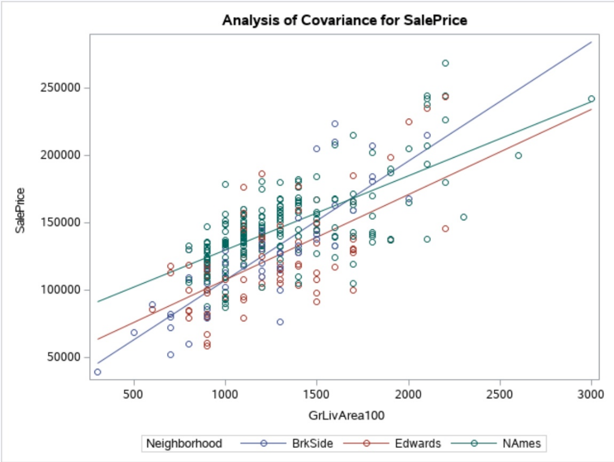

Figure 3.6: Linear Regression Model by Neighborhood.

Root MSE	24094
Dependent Mean	137977
R-Square	0.5333
Adj R-Sq	0.5291
AIC	7037.53791
AICC	7037.72196
BIC	6705.63517
C(p)	4.00000
PRESS	1.961216E11
SBC	6718.75845
ASE	573508240

Table 3.2: Results of Neighborhood Impact on Sales Price

3.4.1. *Adj R^2*

ADJUSTED R^2 obtained for this model is 0.53.

3.4.2. *Internal Press*

PRESS obtained by this model is $1.96E11$.

3.5. Parameters

3.5.1. *Estimates*

The parameter estimates can be seen in Table 3.1. With these estimates a separate equation can be written for each neighborhood to predict sale price based on living area.

$$SalePrice = 74982 + 54.93(GrLivArea) - 55776(BrkSide) - 30274(Edwards) + 33.31(GrLivArea)(BrkSide) + 8.24(GrLivArea)(Edwards)$$

NORTH AMES

$$SalePrice = 74982 + 54.93(GrLivArea)$$

EDWARDS

$$\begin{aligned} SalePrice &= 74982 + 54.93(GrLivArea) - 30274 + 8.24(GrLivArea) \\ &= 44708 + 63.17(GrLivArea) \end{aligned}$$

BROOKSIDE

$$\begin{aligned} SalePrice &= 74982 + 54.93(GrLivArea) - 55776 + 33.31(GrLivArea) \\ &= 19206 + 88.24(GrLivArea) \end{aligned}$$

3.5.2. Interpretation

β_0 The intercept in this model provides an estimate (74982) of the sale price of a home in North Ames (reference neighborhood) with a living area of zero. Of course, this is extrapolation and does not have a clear, practical meaning.

β_1 For each 100 square foot increase in the living area of a home in North Ames, the estimated sale price increases \$54.93.

β_2 This is the adjustment of the intercept for a home in Brookside with respect to a home in North Ames. For a living area of zero, the home in Brookside has an estimated sale price of \$55,776 less than a home in North Ames.

β_3 This is the adjustment of the intercept for a home in Edwards with respect to a home in North Ames. For a living area of zero, the home in Edwards has an estimated sale price of \$30,274 less than a home in North Ames.

β_4 For each 100 square foot increase in the living area of a home in Brookside, the estimated sale price increases \$33 from the change with the home in North Ames.

β_5 For each 100 square foot increase in the living area of a home in Edwards, the estimated sale price increases \$8 from the change with the home in North Ames.

3.5.3. Confidence Intervals

The confidence intervals for the estimates can be seen in Table 3.1.

3.6. Conclusion

The square feet of above ground living area is a statistically significant feature to use to predict the sale price of homes in the North Ames, Edwards, and Brookside neighborhoods of Ames, Iowa. The existing sale prices of homes that have sold in these neighborhoods that are under \$300,000 and underwent a normal sales condition meet all the assumptions required to generate an appropriate linear regression model. We also examined the differences in predicted prices in these three neighborhoods and determined that the sale prices of homes in each of these neighborhoods do differ from each other. Homes in North Ames under 1600 square feet are predicted to sell for the highest price as compared to the other two neighborhoods. The price per square foot in Brookside increases at the highest rate of the three neighborhoods.

Homes greater than approximately 1000 square feet in Brookside are predicted to sell for higher prices than comparable homes in Edwards. Homes greater than approximately 1600 square feet in Brookside are predicted to sell for higher prices than comparable homes in North Ames. Homes in the Edwards neighborhood are predicted to sell for the lowest price of all three neighbor hoods with the exception of homes that are smaller than 1000 square feet.

4. Analysis Question 2

4.1. Restatement of Problem

Build the most predictive model for sales prices of homes in all of Ames Iowa. This includes all neighborhoods. Your group is limited to only the techniques we have learned in 6371 (no random forests or other methods we have not yet covered). Specifically, you should produce 4 models: one from forward selection, one from backwards elimination, one from stepwise selection, and one that you build custom. The custom model could be one of the three preceding models or one that you build by adding or subtracting variables at your will. Generate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model you feel is the best in terms of being able to predict future sale prices of homes in Ames, Iowa.

4.2. Model Selection

4.2.1. Stepwise

4.2.2. Forward

4.2.3. Backward

4.2.4. Custom

4.3. Checking Assumptions

4.3.1. Residual Plots

4.3.2. Influential point analysis (Cook's D and Leverage)

4.3.3. Make sure to address each assumption

4.4. Comparing Competing Models

Predictive Models	**Adjusted R^2**	**CV Press**	**Kaggle Score**
Forward	XX	XX	XX
Backward	XX	XX	XX
Stepwise	XX	XX	XX
CUSTOM	XX	XX	XX

Table 4.1: Analysis Results

4.4.1. Adj R^2

4.4.2. Internal CV Press

4.4.3. Kaggle Score

4.5. Conclusion

A. Source Code for Analysis 1

Listing 1: Analysis 1 SAS Code.

```
*-------------------------------------------------------------*
| Import train.csv                                            |
| Set REFFILE for train.csv                                   |
*-------------------------------------------------------------*;

* FILENAME REFFILE '/home/mwolfe0/train.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

*-------------------------------------------------------------*
| Subset the data to only include homes sold in the           |
| neighborhoods of interest - NAmes, BrkSide, and Edwards|
| Round the gross living area to the nearest 100 SF           |
| Keep only the variables of Neighboorhood, GrLivArea,        |
| and SalePrice in the dataset                                |
*-------------------------------------------------------------*;

DATA HOMES1;
SET TRAIN (KEEP=Neighborhood GrLivArea SalePrice);
IF Neighborhood EQ "NAmes" |
   Neighborhood EQ "BrkSide" |
   Neighborhood EQ "Edwards";
GrLivArea100 = ROUND(GrLivArea, 100); /*FLOOR(GrLivArea);*/
RUN;

*-------------------------------------------------------------*
| Descriptive statistics on the HOMES1 dataset for            |
| GrLivArea and SalePrice                                     |
*-------------------------------------------------------------*;

PROC UNIVARIATE DATA=HOMES1;
        CLASS Neighborhood;
        VAR GrLivArea SalePrice;
RUN;

*-------------------------------------------------------------*
| Scatter plot of sale prices in the three neighborhoods |
| vs Gross Living Area                                        |
*-------------------------------------------------------------*;

PROC SGPLOT DATA=HOMES1;
        SCATTER X=GrLivArea Y=SalePrice;
        REG X=GrLivArea Y=SalePrice;
```

```
RUN;

*------------------------------------------------------------*
| Regression model of homes in the three neighborhoods    |
| combined for Sale Price based on Gross Living Area      |
| to check assumptions on the data in these three         |
| neighborhoods                                            |
*------------------------------------------------------------*;

PROC REG DATA=HOMES1;
        MODEL SalePrice=GrLivArea / CLB;
        RUN;

*------------------------------------------------------------*
| Regression model of homes in the three neighborhoods    |
| using an equal slope model                               |
*------------------------------------------------------------*;

PROC GLM DATA=HOMES1;
        CLASS Neighborhood;
        MODEL SalePrice=GrLivArea Neighborhood / CLPARM;
        RUN;

*------------------------------------------------------------*
| Regression model of homes in the three neighborhoods    |
| using an equal intercept model (slopes differ)          |
*------------------------------------------------------------*;

PROC GLM DATA=HOMES1;
        CLASS Neighborhood;
        MODEL SalePrice=GrLivArea*Neighborhood / CLPARM;
        RUN;

*------------------------------------------------------------*
| Regression model of homes in the three neighborhoods    |
| using a model that allows slopes and intercepts to      |
| vary                                                     |
*------------------------------------------------------------*;

PROC GLM DATA=HOMES1;
        CLASS Neighborhood;
        MODEL SalePrice=Neighborhood GrLivArea*Neighborhood / CLPARM;
        RUN;

*------------------------------------------------------------*
| Remove Outliers:                                         |
| SaleCondition is not normal (confounding effect on      |
| prices)                                                  |
| SalePrice is greater than 300,000 since they are not    |
| representative of the overall population in these      |
```

```
| three neighborhoods.                                  |
| Keep only the variables of Neighboorhood, GrLivArea,  |
| and SalePrice in the dataset                          |
*-------------------------------------------------------*;

DATA HOMES2;
SET TRAIN (KEEP=Neighborhood GrLivArea SalePrice SaleCondition);
IF Neighborhood EQ "NAmes" |
   Neighborhood EQ "BrkSide" |
   Neighborhood EQ "Edwards";
IF SalePrice LT 300000;
IF SaleCondition EQ "Normal";
GrLivArea100 = ROUND(GrLivArea, 100);
RUN;

*-------------------------------------------------------*
| Descriptive statistics on the HOMES1 dataset for      |
| GrLivArea and SalePrice                               |
*-------------------------------------------------------*;

PROC UNIVARIATE DATA=HOMES2;
        CLASS Neighborhood;
        VAR GrLivArea SalePrice;
RUN;

*-------------------------------------------------------*
| Scatter plot of sale prices in the three neighborhoods |
| vs Gross Living Area                                  |
*-------------------------------------------------------*;

PROC SGPLOT DATA=HOMES2;
        SCATTER X=GrLivArea Y=SalePrice;
        REG X=GrLivArea Y=SalePrice;
RUN;

*-------------------------------------------------------*
| Regression model of homes in the three neighborhoods  |
| using a model that allows slopes and intercepts to    |
| vary                                                  |
| Output 95% confidence limit for parameter estimates   |
*-------------------------------------------------------*;

PROC REG DATA=HOMES2 PLOTS=ALL;
        MODEL SalePrice=GrLivArea / CLB;
        RUN;

*-------------------------------------------------------*
| Regression model of homes in the three neighborhoods  |
| using an equal slope model                            |
*-------------------------------------------------------*;
```

```
PROC GLM DATA=HOMES2;
        CLASS Neighborhood;
        MODEL SalePrice=GrLivArea Neighborhood / CLPARM;
        RUN;

*------------------------------------------------------------*
| Regression model of homes in the three neighborhoods      |
| using an equal intercept model (slopes differ)            |
*------------------------------------------------------------*;

PROC GLM DATA=HOMES2;
        CLASS Neighborhood;
        MODEL SalePrice=GrLivArea*Neighborhood / CLPARM;
        RUN;

*------------------------------------------------------------*
| Regression model of homes in the three neighborhoods      |
| using a model that allows slopes and intercepts to        |
| vary                                                      |
*------------------------------------------------------------*;

PROC GLM DATA=HOMES2;
        CLASS Neighborhood;
        MODEL SalePrice=Neighborhood GrLivArea*Neighborhood / CLPARM;
        RUN;


*------------------------------------------------------------*
| Alternate Method with interaction terms                   |
|                                                           |
| Keep only the variables of Neighboorhood, GrLivArea,      |
| and SalePrice in the dataset                              |
| d1 = NAmes, d2 = BrkSide, Control = Edwards               |
*------------------------------------------------------------*;

DATA HOMES3;
SET TRAIN (KEEP=Neighborhood GrLivArea SalePrice SaleCondition);
IF Neighborhood EQ "NAmes" |
   Neighborhood EQ "BrkSide" |
   Neighborhood EQ "Edwards";
IF SalePrice LT 300000;
IF SaleCondition EQ "Normal";
GrLivArea100 = ROUND(GrLivArea, 100);
  IF Neighborhood = 'NAmes' THEN d1 = 1; ELSE d1=0;
  IF Neighborhood = 'BrkSide' THEN d2 = 1; ELSE d2=0;
              int1 = d1*GrLivArea100; int2 = d2*GrLivArea100;
RUN;

*------------------------------------------------------------*
```

```
| Plots to check assumptions                                  |
| d1 = NAmes, d2 = BrkSide, Control = Edwards                 |
*-------------------------------------------------------------*;

PROC SGPLOT DATA=HOMES3;
HISTOGRAM GrLivArea100;
DENSITY GrLivArea100/TYPE=NORMAL;
TITLE "Histogram of Gross Living Area in NAmes, BrkSide, and Edwards";
RUN;

PROC SGPLOT DATA=HOMES3;
SCATTER X=GrLivArea100 Y=SalePrice;
TITLE "Gross Living Area vs Sale Price in NAmes, BrkSide, and Edwards";
RUN;

PROC REG DATA=HOMES3;
model SalePrice = GrLivArea100/CLB;
RUN;

*-------------------------------------------------------------*
| Run regression model with interaction terms using dummy|
| variables                                                   |
| d1 = NAmes, d2 = BrkSide, Control = Edwards                 |
| Output 95% confidence limit for parameter estimates         |
*-------------------------------------------------------------*;
PROC REG DATA=HOMES3;
        model SalePrice = GrLivArea100 d1 d2 int1 int2/VIF CLB;
        title
        'Regression of Sale Price on Gross Living Area
        with Interaction Terms';
        RUN;

*-------------------------------------------------------------*
| center the interaction terms based on the means of          |
| GrLivArea100 and d1 and d2 to correct for the               |
| inflated VIF                                                |
*-------------------------------------------------------------*;

PROC MEANS DATA=HOMES3;
var GrLivArea100 d1 d2;
run;

DATA center;
set Homesp1b;
cent1 = (GrLivArea100 - 1280.72)*(d1-0.588);
cent2 = (GrLivArea100 - 1280.72)*(d2-0.151);
RUN;

DATA center;
set HOMES3;
```

```
cent1 = (GrLivArea100 - 1283.2)*(d1-0.593);
cent2 = (GrLivArea100 - 1283.2)*(d2-0.164);
RUN;

PROC REG DATA=center PLOTS=ALL;
model SalePrice = GrLivArea100 d1 d2 cent1 cent2/VIF CLB;
 title
        'Regression of Sale Price on Gross Living Area
        with Interaction Terms';
RUN;

PROC GLM DATA=HOMES3 PLOT=ALL;
CLASS Neighborhood;
model SalePrice=GrLivArea100|Neighborhood/solution CLPARM;
RUN;
```

B. SOURCE CODE FOR ANALYSIS 2

B.1. Forward Selection

Listing 2: Forward Selection SAS Code.

```
*--------------------------------------------------------------*
| Import train.csv                                              |
| Import test.csv                                               |
| Set REFFILE for train.csv                                     |
| Set REFFILE2 for test.csv                                     |
*--------------------------------------------------------------*;

*FILENAME REFFILE '/home/mwolfe0/train.csv';
*FILENAME REFFILE2 '/home/mwolfe0/test.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';
FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;
        GETNAMES=YES;
RUN;

*--------------------------------------------------------------*
| Combine train and test into one datafile HOMES                |
*--------------------------------------------------------------*;

DATA HOMES;
        SET TRAIN TEST;
        IF LotFrontage EQ "NA" THEN LotFrontage = 0;
        LotFront = input(LotFrontage, 8.);
        drop LotFrontage;
        RENAME LotFront=LotFrontage;
RUN;

*--------------------------------------------------------------*
| Code for forward selection                                    |
| Set seed to a constant for model comparison                   |
| Class variable input with split option to allow               |
|       classification variable to be able to enter or          |
|       leave the model independently                           |
| Stop=CV specifies the model will stop when the                |
|       predicted residual sum of square is reached with        |
|       k-fold cross validation                                 |
| CVMethod specifies how subsets ar formed for                  |
|       cross validation                                        |
```

```
| OUTPUT Dataset to RESULTS with the predicted variable    |
|         based on the final model                         |
*----------------------------------------------------------*;

PROC GLMSELECT DATA=HOMES SEED=71669132;
        CLASS MSSubClass MSZoning Street Alley LotShape LandContour
                   Utilities LotConfig LandSlope Neighborhood Condition1
                   Condition2 BldgType HouseStyle OverallQual OverallCond
                   RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                   ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                   BsmtFinType2 Heating HeatingQC CentralAir Electrical
                   KitchenQual Functional FireplaceQu GarageType
                   GarageFinish GarageQual GarageCond PavedDrive PoolQC
                   Fence MiscFeature SaleType SaleCondition RoofStyle
                   BsmtCond MasVnrType
                   / split;
        MODEL SalePrice= LotArea YearBuilt YearRemodAdd BsmtFinSF1
                   BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF
                   LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
                   FullBath HalfBath BedroomAbvGr KitchenAbvGr
                   TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
                   GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
                   _3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold
                   MSSubClass MSZoning Street Alley LotShape LandContour
                   Utilities LotConfig LandSlope Neighborhood Condition1
                   Condition2 BldgType HouseStyle OverallQual OverallCond
                   RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                   ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                   BsmtFinType2 Heating HeatingQC CentralAir Electrical
                   KitchenQual Functional FireplaceQu GarageType
                   GarageFinish GarageQual GarageCond PavedDrive PoolQC
                   Fence MiscFeature SaleType SaleCondition LotFrontage
                   RoofStyle BsmtCond MasVnrType
        / selection =forward(stop=CV) cvmethod=random(5) stats=all;
        OUTPUT OUT=RESULTS P=PREDICT;
RUN;

*----------------------------------------------------------*
| Create a datafile RESULTS_FW of predicted values for     |
| SalePrice for house id greater than 1460 which           |
| is where the Kaggle test set data begins.                |
*----------------------------------------------------------*;

DATA RESULTS_FW;
        SET RESULTS;

        IF PREDICT > 0 THEN
                SalePrice=Predict;

        IF PREDICT < 0 THEN
```

```
                SalePrice=10000;
        KEEP id SalePrice;
        WHERE id > 1460;
RUN;

*------------------------------------------------------------*
| Export a datafile for predicted values for                 |
| SalePrice for house id greater than 1460 which             |
| is where the Kaggle test set data begins.                  |
*------------------------------------------------------------*;

*FILENAME REFFILE3 '/home/mwolfe0/results_fw.csv';
FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_fw.csv';

PROC EXPORT DATA=RESULTS_FW FILE=REFFILE3 DBMS=CSV REPLACE;
RUN;
```

B.2. Backward Selection

Listing 3: Backward Selection SAS Code.

```
*------------------------------------------------------------*
| Import train.csv                                           |
| Import test.csv                                            |
| Set REFFILE for train.csv                                  |
| Set REFFILE2 for test.csv                                  |
*------------------------------------------------------------*;

*FILENAME REFFILE '/home/mwolfe0/train.csv';
*FILENAME REFFILE2 '/home/mwolfe0/test.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';
FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;
        GETNAMES=YES;
RUN;

*------------------------------------------------------------*
| Combine train and test into one datafile HOMES             |
*------------------------------------------------------------*;

DATA HOMES;
        SET TRAIN TEST;
```

```
        IF LotFrontage EQ "NA" THEN LotFrontage = 0;
        LotFront = input(LotFrontage, 8.);
        drop LotFrontage;
        RENAME LotFront=LotFrontage;
RUN;

*--------------------------------------------------------------------*
| Code for backward selection                                        |
| Set seed to a constant for model comparison                        |
| Class variable input with split option to allow                    |
|       classification variable to be able to enter or               |
|       leave the model independently                                |
| Stop=10 specifies the model will stop selection at the             |
|       first step for which the selected model has 10               |
|       effects                                                      |
| CVMethod specifies how subsets ar formed for                       |
|       cross validation                                             |
| OUTPUT Dataset to RESULTS with the predicted variable              |
|        based on the final model                                    |
*--------------------------------------------------------------------*;

PROC GLMSELECT DATA=HOMES SEED=71669132;
        CLASS MSSubClass MSZoning Street Alley LotShape LandContour
                    Utilities LotConfig LandSlope Neighborhood Condition1
                    Condition2 BldgType HouseStyle OverallQual OverallCond
                    RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                    ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                    BsmtFinType2 Heating HeatingQC CentralAir Electrical
                    KitchenQual Functional FireplaceQu GarageType
                    GarageFinish GarageQual GarageCond PavedDrive PoolQC
                    Fence MiscFeature SaleType SaleCondition RoofStyle
                    BsmtCond MasVnrType
                    / split;
        MODEL SalePrice= LotArea YearBuilt YearRemodAdd BsmtFinSF1
                    BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF
                    LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
                    FullBath HalfBath BedroomAbvGr KitchenAbvGr
                    TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
                    GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
                    _3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold
                    MSSubClass MSZoning Street Alley LotShape LandContour
                    Utilities LotConfig LandSlope Neighborhood Condition1
                    Condition2 BldgType HouseStyle OverallQual OverallCond
                    RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                    ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                    BsmtFinType2 Heating HeatingQC CentralAir Electrical
                    KitchenQual Functional FireplaceQu GarageType
                    GarageFinish GarageQual GarageCond PavedDrive PoolQC
                    Fence MiscFeature SaleType SaleCondition LotFrontage
                    RoofStyle BsmtCond MasVnrType
```

```
        / selection =backward(stop=10) cvmethod=random(5)
                    stats=ADJRSQ stats=PRESS;
        OUTPUT OUT=RESULTS P=PREDICT;
RUN;

*-------------------------------------------------------*
| Create a datafile RESULTS_BW of predicted values for  |
| SalePrice for house id greater than 1460 which        |
| is where the Kaggle test set data begins.             |
*-------------------------------------------------------*;

DATA RESULTS_BW;
        SET RESULTS;

        IF PREDICT > 0 THEN
                SalePrice=Predict;

        IF PREDICT < 0 THEN
                SalePrice=10000;
        KEEP id SalePrice;
        WHERE id > 1460;
RUN;

*-------------------------------------------------------*
| Export a datafile for predicted values for            |
| SalePrice for house id greater than 1460 which        |
| is where the Kaggle test set data begins.             |
*-------------------------------------------------------*;

*FILENAME REFFILE3 '/home/mwolfe0/results_bw.csv';
FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_bw.csv';

PROC EXPORT DATA=RESULTS_BW FILE=REFFILE3 DBMS=CSV REPLACE;
RUN;
```

B.3. Stepwise Selection

Listing 4: Stepwise Selection SAS Code.

```
*-------------------------------------------------------*
| Import train.csv                                      |
| Import test.csv                                       |
| Set REFFILE for train.csv                             |
| Set REFFILE2 for test.csv                             |
*-------------------------------------------------------*;

*FILENAME REFFILE '/home/mwolfe0/train.csv';
*FILENAME REFFILE2 '/home/mwolfe0/test.csv';
FILENAME REFFILE
```

```
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';
FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;
        GETNAMES=YES;
RUN;

*------------------------------------------------------------------*
| Combine train and test into one datafile HOMES                   |
*------------------------------------------------------------------*;

DATA HOMES;
        SET TRAIN TEST;
        IF LotFrontage EQ "NA" THEN LotFrontage = 0;
        LotFront = input(LotFrontage, 8.);
        drop LotFrontage;
        RENAME LotFront=LotFrontage;
RUN;

*------------------------------------------------------------------*
| Code for stepwise selection                                      |
| Set seed to a constant for model comparison                      |
| Class variable input with split option to allow                  |
|        classification variable to be able to enter or            |
|        leave the model independently                             |
| Stop=CV specifies the model will stop when the                   |
|        predicted residual sum of square is reached with          |
|        k-fold cross validation                                   |
| CVMethod specifies how subsets ar formed for                     |
|        cross validation                                          |
| OUTPUT Dataset to RESULTS with the predicted variable            |
|         based on the final model                                 |
*------------------------------------------------------------------*;

PROC GLMSELECT DATA=HOMES SEED=71669132;
        CLASS MSSubClass MSZoning Street Alley LotShape LandContour
                Utilities LotConfig LandSlope Neighborhood Condition1
                Condition2 BldgType HouseStyle OverallQual OverallCond
                RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                BsmtFinType2 Heating HeatingQC CentralAir Electrical
                KitchenQual Functional FireplaceQu GarageType
                GarageFinish GarageQual GarageCond PavedDrive PoolQC
                Fence MiscFeature SaleType SaleCondition RoofStyle
                BsmtCond MasVnrType
```

```
                / split;
        MODEL SalePrice= LotArea YearBuilt YearRemodAdd BsmtFinSF1
                BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF
                LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
                FullBath HalfBath BedroomAbvGr KitchenAbvGr
                TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
                GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
                _3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold
                MSSubClass MSZoning Street Alley LotShape LandContour
                Utilities LotConfig LandSlope Neighborhood Condition1
                Condition2 BldgType HouseStyle OverallQual OverallCond
                RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                BsmtFinType2 Heating HeatingQC CentralAir Electrical
                KitchenQual Functional FireplaceQu GarageType
                GarageFinish GarageQual GarageCond PavedDrive PoolQC
                Fence MiscFeature SaleType SaleCondition LotFrontage
                RoofStyle BsmtCond MasVnrType
        / selection=stepwise(stop=CV) cvmethod=random(5) stats=all;
        OUTPUT OUT=RESULTS P=PREDICT;
RUN;

*-----------------------------------------------------------*
| Create a datafile RESULTS_SW of predicted values for     |
| SalePrice for house id greater than 1460 which           |
| is where the Kaggle test set data begins.                |
*-----------------------------------------------------------*;

DATA RESULTS_SW;
        SET RESULTS;

        IF PREDICT > 0 THEN
                SalePrice=Predict;

        IF PREDICT < 0 THEN
                SalePrice=10000;
        KEEP id SalePrice;
        WHERE id > 1460;
RUN;

*-----------------------------------------------------------*
| Export a datafile for predicted values for               |
| SalePrice for house id greater than 1460 which           |
| is where the Kaggle test set data begins.                |
*-----------------------------------------------------------*;

*FILENAME REFFILE3 '/home/mwolfe0/results_sw.csv';
FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_sw.csv';

```

```
PROC EXPORT DATA=RESULTS_SW FILE=REFFILE3 DBMS=CSV REPLACE;
RUN;
```

C. High Level Summary of Data

Attribute Type		Description	Features
Categorical (Qualitative)	Nominal	Only provide enough information to distinguish one object from another	MSSubClass, MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, Foundation, BsmtExposure, Heating CentralAir, Electrical, GarageType, GarageFinish PavedDrive, MiscFeature, SaleType, SaleCondition RoofStyle **(30 variables)**
	Ordinal	Provide enough information to order objects	OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, BsmtFinType1, BsmtFinType2, HeatingQC, KitchenQual, Functional, FireplaceQu, GarageQual, GarageCond, PoolQC, Fence **(16 variables)**
Numeric (Quantitative)	Interval	Interval attributes difference between values are meaningful	YearBuilt, YearRemodAdd, GarageYrBlt, MoSold, YrSold **(5 variables)**
	Ratio	Differences and ratios are meaningful	LotFrontage, LotArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, MasVnrArea, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal **(28 variables)**

Table C.1: Summary of Dataset

D. DETAILED DATA DESCRIPTION

Classification Variables		
MSSubClass	Identifies the type of dwelling involved in the sale	
	20	1-STORY 1946 & NEWER ALL STYLES
	30	1-STORY 1945 & OLDER
	40	1-STORY W/FINISHED ATTIC ALL AGES
	45	1-1/2 STORY - UNFINISHED ALL AGES
	50	1-1/2 STORY FINISHED ALL AGES
	60	2-STORY 1946 & NEWER
	70	2-STORY 1945 & OLDER
	75	2-1/2 STORY ALL AGES
	80	SPLIT OR MULTI-LEVEL
	85	SPLIT FOYER
	90	DUPLEX - ALL STYLES AND AGES
	120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
	150	1-1/2 STORY PUD - ALL AGES
	160	2-STORY PUD - 1946 & NEWER
	180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
	190	2 FAMILY CONVERSION - ALL STYLES AND AGES