Century 21 Housing Sales Report 2022

Ames, Iowa

Miguel Bonilla, Laura Ahumada, Zack Bunn

Data Description:

With respect to our client, Century 21, we were tasked with investigating Housing sales data for neighborhoods in Ames, Iowa. This data covers a diverse set of variables that explain each feature and amenity that these homes offer. With 79 columns and 1456 observations we decided to go with what we found to be key variables in our study to find the newest trends in the market. Notable variables include sale price, above ground living area (GrLivArea), neighborhood, lot area, first floor square footage (1stFlrSf), second floor square footage (2ndFlrSf), and masonry veneer area (MasVnrArea), among others.

Analysis 1:

Century 21 has tasked us with analyzing the sales of North Ames, Edwards and Brooke Side, the three neighborhoods they sell homes in. For this study we will find an estimate for how the Sales prices are related to the Square Footage. In other words, we would like to find out if the median sale price changes as the above ground living area increases. (Code starts on page 8)

Assumptions:

Logarithmic transformations were applied to both sale price and above ground living area, which resulted in an approximately linear scatterplot (**Figure 1**). From our exploratory data analysis, we discovered there were observations with high residuals and high leverage (**Appendix 1**). Looking at the Cook's D and the Studentized Residuals, we identified a particularly influential observation. When digging deeper through our dataset, we noticed this observation corresponded to a home with over 5000 sqft of above ground living area sold for a much lower price than expected for a size that size ($160,000), while we are not able to spend the time and money determining if this was a recording error we decided to focus our further study on homes below 4000 sqft of above ground construction since homes of that size are very rare in the market and not particularly useful in predicting future sales for the company.

After filtering out homes of over 4000 sqft, we can see the resulting plots show normal distributions for homes of a given size, and with approximately equal spread (standard deviations) (**Appendix 2**). There seems to be an outlier per the Leverage vs Rstudent plot but that high cook's D point is only

0.07 which is not very significant so we can dismiss it. There is linearity between the response variable and the predictor, and the normality assumptions are met so we now can perform linear regression.

Figure 1 Log Sale Price vs. Log Abv. Gr. Living Area

Comparing Competing Models

```
/* run selection procedure with and without interactions to check models with these variables */
proc glmselect data=q1_filtered;
class neighborhood;
model lsaleprice =lgrlivarea neighborhood/showpvalues selection=forward(include=2 choose=cv);
run;
proc glmselect data = q1_filtered;
class neighborhood;
model lsaleprice =lgrlivarea|neighborhood/ showpvalues selection=forward(include=3 choose=cv);
run;
```

Figure 2 SAS code for comparing models without and with interactions

We fitted two separate models for comparison, one using a log of above ground living area and neighborhood to predict log sale price, and a separate model including interactions between neighborhood and the log of the area. Both resulted in models that were statistically significant (both P-values<.0001), we compared the models using different goodness-of-fit metrics and the model with interactions had both a larger Adjusted R-squared of .5216 and smaller CV Press of 14.124 (**Table 1**). Additionally, we conducted an analysis of variance test (ANOVA) to check if there was a significant difference between the two models (**Table 2**).

	Model without Interactions	Model with Interactions
Adj. R-sq	0.5002	0.5216
CV PRESS	14.40244	14.34441

Table 1 Comparison of models with and without interactions

Source	DF	Sum of Squares	Mean Square	F-Value	Pr>F
Model	2	0.67618	0.33809	9.449134	<.0001
Error	375	13.41833	0.03578		
Total	377	14.09451			

Table 2 ANOVA table for comparing models without (Total) and with interactions (Error)

The ANOVA test produced evidence that the two models are statistically different (p<.0001), this information, coupled with the fact that the model has a higher R-squared and lower CV PRESS value leads us choose the model with interactions between the log of above ground living area and Neighborhoods (Edwards, North Ames, and Brookside).

All of the parameter estimates on our model were statistically significant at the .05 level (**Figure 3**). The resulting model is included below.

```
proc glm data = q1_filtered plots=all;
class Neighborhood;
model lSalePrice = lGrLivArea Neighborhood/solution;
run;
```

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	8.492727641	B	0.31916233	26.61	<.0001	7.865155507	9.120299776
lgrlivarea	0.473023602	B	0.04469311	10.58	<.0001	0.385143083	0.560904121
Neighborhood Brk Side	-2.579806905	B	0.59016472	-4.37	<.0001	-3.740253790	-1.419360020
Neighborhood Edwards	-1.569634752	B	0.58560055	-2.68	0.0077	-2.721107064	-0.418162440
Neighborhood NAmes	0.000000000	B	.	.	.	.	.
lgrlivare*Neighborho Brk Side	0.346624454	B	0.08344620	4.15	<.0001	0.182543340	0.510705568
lgrlivare*Neighborho Edwards	0.200314013	B	0.08227532	2.43	0.0154	0.038535213	0.362092813
lgrlivare*Neighborho NAmes	0.000000000	B	.	.	.	.	.

Figure 3 Parameter estimates for model with interactions between lGrLivArea and Neighborhood

$$Pred.log\ Sale\ Price\ = 8.49 + .473\ lGrLivArea + .347 BrkSide * lGrLivArea + .200 Edwards * lGrLivArea$$

Summary of the analysis:

Since we ran the analysis for the three different neighborhoods, we will look at the results for each individual area.

North Ames:

$$\{lSalePrice|lGrLivArea\ Neighborhood = NAmes\} = 8.493 + .473lGrLivArea$$

For the North Ames neighborhood, a doubling of above ground living area is associated with a 38.8% increase in median sale price (P-value <.001).. Moreover, we are 95% confident that for the North Ames neighborhood, this median sales price increase is between $(1.306, 1.475) = 31\%\ to\ 48\%$.

Brookside:

$$\{lSalePrice|lGrLivArea\ Neighborhood = Brkside\} = 5.912 + .81lGrLivArea$$

For the Brookside neighborhood, a doubling of above ground living area is associated with a 76.4% increase in median sale price (P-value<.001). Moreover, we are 95% confident that this median price increase is between $(1.48, 2.11) = 48.2\%\ to\ 111\%$.

Edwards:

$$\{lSalePrice|lGrLivArea\ Neighborhood = Edwards\} = 6.923 + .673lGrLivArea$$

For the Edwards Neighborhood, a doubling of above ground living area is associated with a 59.4% increase in median sale price (P-value<.001). Moreover, we are 95% confident that this median sale price increase is between $(1.341,1.896) = 34.1\%\ to\ 89.6\%$.

Analysis 2:

To get a greater understanding of the Housing market of Ames as a whole, we will analyze sales data across the whole city. Our goal here is to find an accurate predictive model that will give Century 21 a competitive edge in this emergent market. To do so we will be comparing different selection methods, Stepwise, Forward, Backward, and our own custom model.

Similar to Analysis one, we will limit the data used to train the model to homes with less than 4000 sqft. of above ground living area, since there are very few houses in the city which are that large, and we think it makes more sense to focus our efforts on houses that are more representative Ames, Iowa.

Additionally, we created additional data features based on already included information. We determined the actual age of homes at time of sale based on the built and sale years, we also created binomial categories for the presence of basement (bsmt), second floor (scndflr), and masonry veneer (vnr).

Assumptions

The residual plots for all models, Forward Selection, Backward selection, Stepwise Selection and custom model, are well distributed, the QQ plots shows slight deviation from normality but due to sample size we can call CLT (**Figures 4-7)**. The residual plots of the response variable versus the predicted values shows that the variance is equal. The leverage vs r student does show that there are influential values but checking the cook's D, the highest residual is very low, 0.05 so we can continue. All explanatory variables are independent and they do have a linear relationship with the response variable (**Appendix 3**). The models are robust for linear regression. regression.

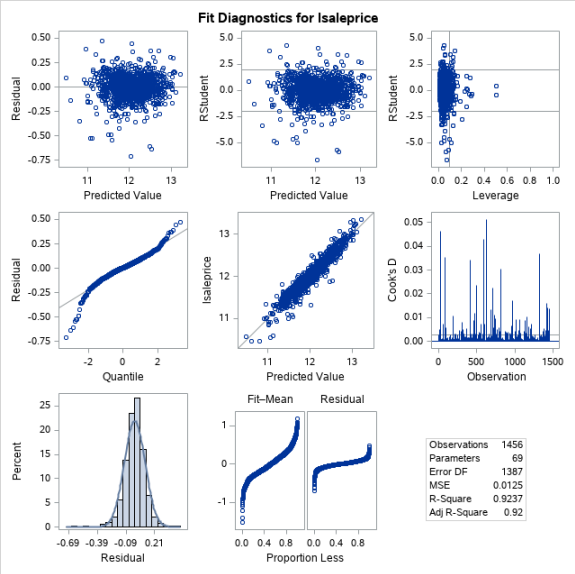


Figure 4 Fit Plot for model from Forward Selection

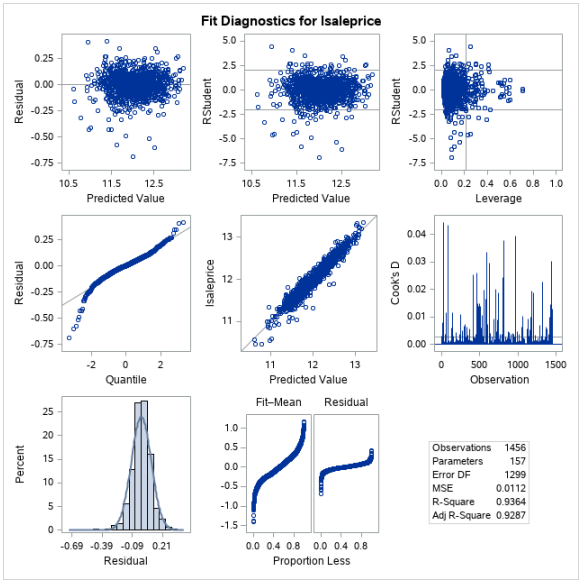


Figure 5 Fit Plot for model from Backward Selection

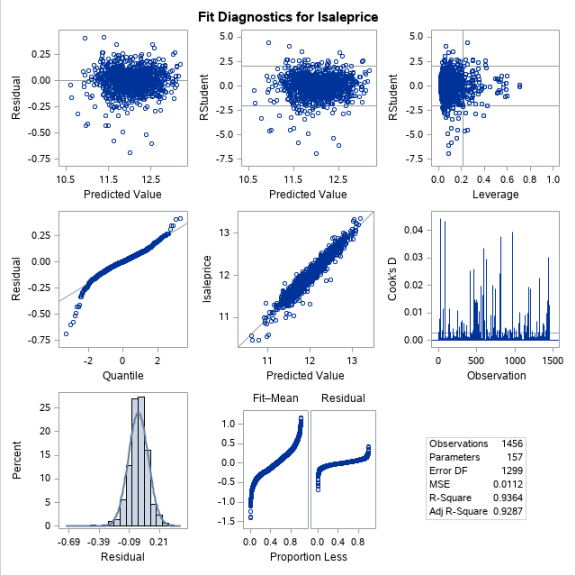


Figure 6 Fit Plot for model from Stepwise Selection

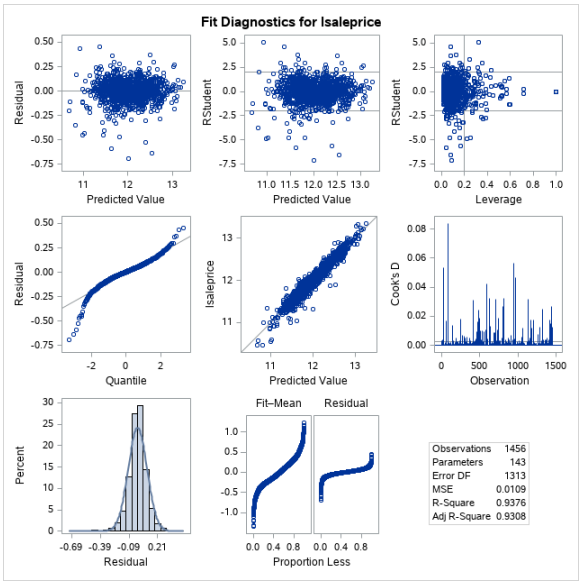


Figure 7 Fit Plot for custom model

Comparing Competing Models:

Based on our results we concluded that the backwards automatic selection method produced the best results given that it resulted in both a higher R-square (.9287) and lower CV PRESS (19.64581) values (**Table 3**). We then fitted a custom model using our knowledge of variables which we deemed significant such as the presence and size of a second floor, the presence and type of masonry veneer front, etc. The SAS code used to select and fit the models is included in the Appendix.

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Forward Selection	**0.9200**	**19.74531**	**0.13498**
Backward Selection	**0.9287**	**19.64581**	**0.13465**
Stepwise Selection	**0.9259**	**18.81634**	**0.13509**
CUSTOM Model	**0.9308**	**19.66267**	**0.13404**

Table 3 Selection Method comparison

For reproducibility of our findings, we have included the selected models for each one of the methods (**Appendix 4-7**), as well as the code for the entire analysis (**Appendix 9**)

Conclusion:

The regression model was successful at predicting median Sale Price. The most successful model produced was the custom-built model (Kaggle score .13404). The key variables found in the analysis were the living area, lot area, size of the first floor, bedrooms above ground, month sold, overall condition, overall quality, year built, year remodeled, and year sold. We were able to create a prediction that is at the bellow the 30th percentile in accuracy in Kaggle using the Custom Model.

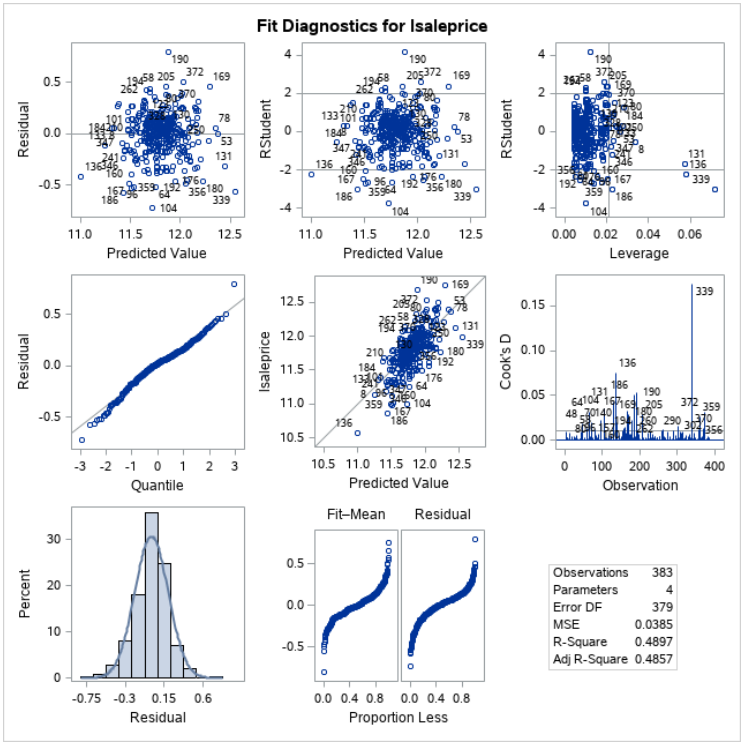
The custom model was statistically significant, with an F-statistic of 138.89, and a p-value<.0001 (**Table 4**). Moreover, this model resulted in the most accurate predicted sale prices for the test set, based on the Kaggle score. The statistical significance of the model, coupled with the higher Kaggle score, gives us confidence that the custom-built model performs better than the ones generated from automatic selection procedures.

Source	DF	Sum of Squares	Mean Square	F-Value	Pr>F
Model	142	214.0112334	1.5071214	138.89	<.0001
Error	1313	14.2480671	0.0108515		
Total	1455	228.2593005			

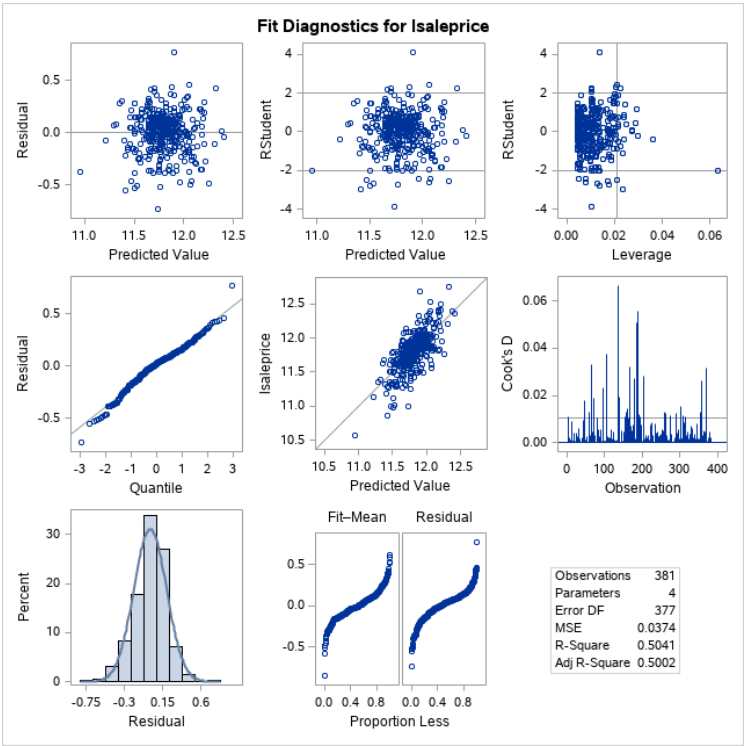
Table 4 Linear Regression ANOVA for Custom Model

If you have any questions, please contact Laura Ahumada (lahumada@mail.smu.edu), Miguel Bonilla (mbonilla@mail.smu.edu), or Zack Bunn (zackb@mail.smu.edu).

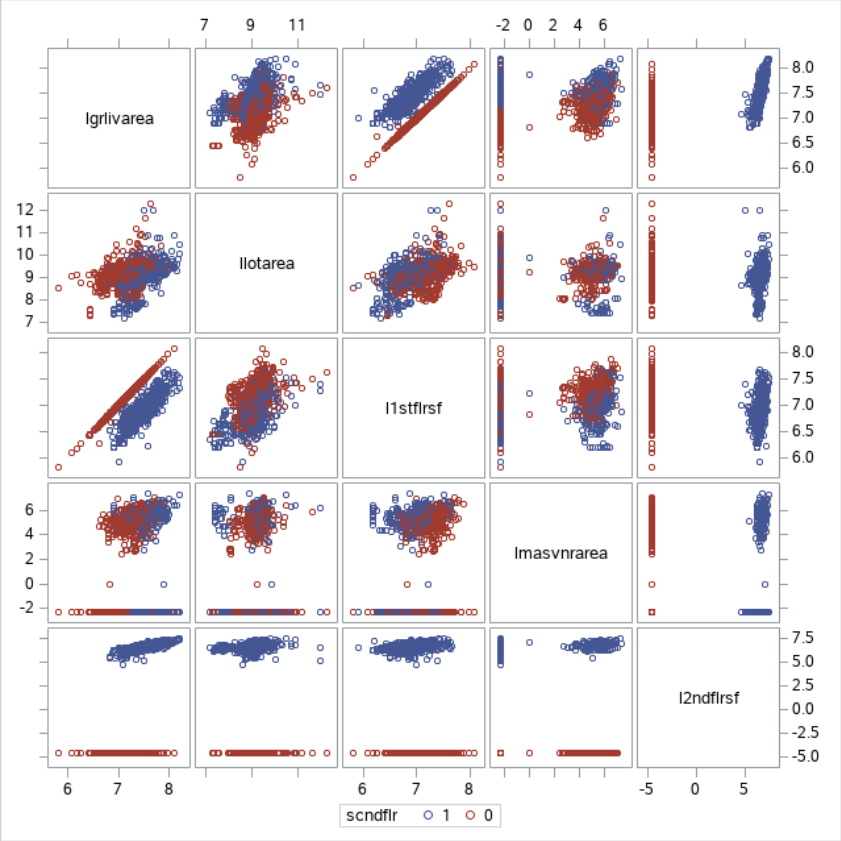
Appendix



Appendix 1 Fit Plots for Log Sale Price vs Log AbvGrLivArea (BrkSide, Edwards, NAmes)



Appendix 2 Fit Plots for Log Sale Price vs Log AbvGrLivArea<4000 (BrkSide, Edwards, NAmes)



Appendix 3 Scatterplot Matrix for Analysis 2

```
lsaleprice=LGRLIVAREA+LLOTAREA+L1STFLRSF+BEDROOMABVGR+OVERALLCOND+
OVERALLQUAL+YEARBUILT+YEARREMODADD+YRSOLD+MSZONING
NEIGHBORHOOD+BSMTQUAL+BSMTEXPOSURE+BSMTFINTYPE1+CENTRALAIR+KITC
HENQUAL+FUNCTIONAL+GARAGEFINISH+SALECONDITION
```

Appendix 4 Forward Selected Model

```
lsaleprice=LGRLIVAREA+LLOTAREA+L1STFLRSF+BEDROOMABVGR+MOSOLD+OVERA
LLCOND+OVERALLQUAL+YEARBUILT+YEARREMODADD+YRSOLD
MSZONING+STREET+ALLEY+LOTSHAPE+LANDCONTOUR+UTILITIES+LOTCONFIG+LA
NDSLOPE+NEIGHBORHOOD+CONDITION1+BLDGTYPE+MASVNRTYPE+EXTERQUAL
EXTERCOND+FOUNDATION+BSMTQUAL+BSMTCOND+BSMTEXPOSURE+BSMTFINTY
PE1+HEATING+HEATINGQC+CENTRALAIR+KITCHENQUAL+FUNCTIONAL
FIREPLACEQU+GARAGETYPE+GARAGEFINISH+GARAGEQUAL+GARAGECOND+PAVE
DDRIVE+POOLQC+FENCE+MISCFEATURE+SALECONDITION
```

Appendix 5 Backward Selected Model

```
lsaleprice=LGRLIVAREA+LLOTAREA+L1STFLRSF+BEDROOMABVGR+OVERALLCOND+
OVERALLQUAL+YEARBUILT+YEARREMODADD+MSZONING
LOTCONFIG+LANDSLOPE+NEIGHBORHOOD+CONDITION1+MASVNRTYPE+BSMTQUA
L+BSMTEXPOSURE+BSMTFINTYPE1+HEATING+HEATINGQC+CENTRALAIR
KITCHENQUAL+FUNCTIONAL+GARAGETYPE+GARAGEFINISH+GARAGECOND+POOL
QC+SALECONDITION
```

Appendix 6 Stepwise Selected Model

```
Lsaleprice =
LGRLIVAREA+LLOTAREA+L1STFLRSF+BEDROOMABVGR+MOSOLD+OVERALLCOND+
OVERALLQUAL
YEARBUILT+YEARREMODADD+YRSOLD+MSZONING+STREET+ALLEY+LANDCONTOU
R+UTILITIES+LOTCONFIG
LANDSLOPE+NEIGHBORHOOD+CONDITION1+BLDGTYPE+EXTERQUAL+EXTERCOND
+FOUNDATION+BSMTQUAL
BSMTCOND+BSMTEXPOSURE+BSMTFINTYPE1+HEATING+HEATINGQC+CENTRALAIR
+KITCHENQUAL+FUNCTIONAL
GARAGEFINISH+GARAGEQUAL+GARAGECOND+PAVEDDRIVE+POOLQC+SALECONDI
TION
AGE+SCNDFLR+AGE*KITCHENQUAL+SCNDFLR*L1STFLRSF+LLOTAREA*L1STFLRSF+
L2NDFLRSF*SCNDFLR+SCNDFLR*BEDROOMABVGR
VNR*LMASVNRAREA+BSMT+BSMT*TOTALBSMTSF+BSMT*BSMTFINSF2
```

Appendix 7 Custom Model

```
/*save file as work.import;/

/*log-log*/
/* add log of sale price and living area */
data train; */
set work.import;
lSalePrice = log(SalePrice);
lGrLivArea = log(GrLivArea);
run;

/* filter 3 neighborhoods for question 1 */
data q1_unfiltered;
set data_Train;
where Neighborhood in ('Edwards', 'NAmes', 'BrkSide');
Run;

/*Check lGrLivArea*/
proc univariate data = q1_unfiltered;
var lGrLivArea;
histogram lGrLivArea;
qqplot lGrLivArea;
Run;

/*Check linear relationship between sales and living area;
proc sgscatter data=q1_unfiltered;
plot lSalePrice*lGrLivArea;
title "lSalePrice vs lGrLivArea";
Run;

/*  run model */
/* observation 339 high leverage high residual cook's d of 2.5 (over 1) is influential */
proc glm data=q1_unfiltered plots=diagnostics(label);
class Neighborhood;
model SalePrice = GrLivArea Neighborhood;
run;

/* filter outliers with area < 4000 */
/*new df called q1_filtered*/
data q1_filtered;
set q1_unfiltered;
where lGrLivArea <400;
run;

/* View updated data */
proc print data=q1_filtered;
run;

/* on filtered data grlivarea<4000sqft */
```

```
/* influential points have been handled, no observations with concerning high leverage high residual,
cook's d's are smaller (under 1) */
/* no evidence against normality or linearity assumptions */
proc glm data = q1_filtered plots=diagnostics(label);
class Neighborhood;
model lSalePrice = lGrLivArea Neighborhood;
run;

/*filtered no interaction model plots */
proc glm data = q1_filtered plots=all;
class Neighborhood;
model lSalePrice = lGrLivArea Neighborhood/solution;
run;
*0.5 r^2;
*cooksD points are not high once the data is filtered to <400 sq ft.

/* filtered interaction model plots*/
proc glm data = q1_filtered plots=all;
class Neighborhood;
model lSalePrice = lGrLivArea| Neighborhood/solution;
run;
*0.52 r^2;

/* run selection procedure with and without interactions to check models */
proc glmselect data=q1_filtered;
class Neighborhood;
model lSalePrice = lGrLivArea Neighborhood/showpvalues selection=forward(include=2 choose=cv);
run;
/* Root MSE           0.19335 */
/* Dependent Mean 11.79752 */
/* R-Square           0.5041 */
/* Adj R-Sq           0.5002 */
/* AIC       -865.16241 */
/* AICC      -865.00241 */
/* SBC       -1232.39121 */
/* CV PRESS           14.99810 */

/*Best model was with the interaction*/
proc glmselect data = q1_filtered;
class Neighborhood;
model lSalePrice = lGrLivArea| Neighborhood/ showpvalues selection=forward(include=3 choose=cv);
run;
/* Root MSE           0.18916 */
/* Dependent Mean 11.79752 */
/* R-Square           0.5279 */
/* Adj R-Sq           0.5216 */
/* AIC       -879.89375 */
/* AICC      -879.59349 */
/* SBC       -1239.23696 */
/* CV PRESS           14.12428 */

/* interactions model seems to be better based on lower cvpress and r-squared,
so it seems the extra predictive power from the additional estimates offsets the loss of degrees of freedom for the
standard errors*/

/*Confidence intervals using proc glm */
proc glm data=q1_filtered;
class Neighborhood;
model lSalePrice = lGrLivArea| Neighborhood/solution clparm;
Run;
```

Appendix 8 SAS code Analysis 1

```
/* Generated Code (IMPORT) */
/* Source File: train.csv */
/* Source Path: /home/u60173286 */
/* Code generated on: 4/5/22, 10:35 PM */

%web_drop_table(WORK.clean);
```

```
FILENAME REFFILE '/home/u60173286/train.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.train;
	GETNAMES=YES;
	guessingrows=1461;
RUN;

PROC CONTENTS DATA=WORK.train; RUN;


%web_open_table(WORK.clean);

/*clean train data */
data trainclean;
set work.train;
lsaleprice = log(saleprice); /*log transform saleprice and areas */
lgrlivarea = log(grlivarea);
llotarea = log(lotarea);
l1stflrsf = log('1stFlrSF'n);
if '2ndflrsf'n>0 then l2ndflrsf=log('2ndflrsf'n); else l2ndflrsf=log(.01);
where grlivarea<4000;
age=YrSold-YearBuilt; /* create time variables for house age and remodel age */
remodage=YrSold-YearRemodAdd;
relage = age-(age-remodage)*1/2; /*relative age variable giving remodel half the weight of build age */
if '2ndflrsf'n =0 then scndflr=0;else scndflr=1; /*dummy variables conditional on presence of feature */
if masvnrarea =0 then vnr=0; else vnr=1;
if masvnrarea >0 then lmasvnrarea=log(masvnrarea); else lmasvnrarea=log(.1);
if garagearea=0 then garage=0; else garage=1;
if garagearea>0 then lgaragearea=log(garagearea); else lgaragearea=log(.1);
if totalbsmtsf=0 then bsmt=0; else bsmt=1;
run;

/* Generated Code (IMPORT) */
/* Source File: test.csv */
/* Source Path: /home/u60173286 */
/* Code generated on: 4/6/22, 10:16 AM */

%web_drop_table(WORK.test);


FILENAME REFFILE '/home/u60173286/test.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.test;
	GETNAMES=YES;
	guessingrows=1460;
RUN;

PROC CONTENTS DATA=WORK.test; RUN;


%web_open_table(WORK.test);

/* clean test */
data test2;
set work.test;
lgrlivarea = log(grlivarea);
llotarea = log(lotarea);
l1stflrsf = log('1stFlrSF'n);
kitchenqual = 'TA';
if kitchenqual = 'NA' then kitchenqual = 'TA'; /*extra levels on test set not present on train, impute mode of test*/
if exterior1st = 'NA' then exterior1st = 'VinylSd';
if exterior2nd  = 'NA' then exterior2nd = 'VinylSd';
if functional = 'NA' then functional = 'Typ';
if garagecars = 'NA' then garagecars = 2;
if totalbsmtsf = 'NA' then totalbsmtsf=0;
```

```
if BsmtUnfSf = 'NA' then BsmtUnfSf=0;
if BsmtFinSf2 = 'NA' then BsmtFinSf2=0;
garagenum=input(garagecars, 8.); /*garage cars read as cat. variable, changing to num. */
drop garagecars;
rename garagenum=GarageCars;
garagecars = input(garagecars, 8.);
if mszoning = 'NA' then mszoning = 'RL';
if saletype = 'NA' then saletype ='WD';
if utilities = 'NA' then utilities = 'AllPub';
format _character_;
age=YrSold-YearBuilt;
remodage=YrSold-YearRemodAdd;
relage = age-(age-remodage)*1/2;
if '2ndflrsf'n =0 then scndflr=0;else scndflr=1;
if '2ndflrsf'n>0 then l2ndflrsf=log('2ndflrsf'n); else l2ndflrsf=log(.01);
if masvnrarea =0 then vnr=0; else vnr=1;
if masvnrarea >0 then lmasvnrarea=log(masvnrarea); else lmasvnrarea=log(.1);
if garagearea=0 then garage=0; else garage=1;
if garagearea>0 then lgaragearea=log(garagearea); else lgaragearea=log(.1);
if totalbsmtsf=0 then bsmt=0; else bsmt=1;
bsmtnum=input(totalbsmtsf, 8.);
drop totalbsmtsf;
rename bsmtnum=TotalBsmtSF;
bsmtunf=input(bsmtunfsf, 8.);
drop bsmtunfsf;
rename bsmtunf=BsmtUnfSf;
bsmtfin2num=input(bsmtfinsf2, 8.);
drop bsmtfinsf2;
rename bsmtfin2num=BsmtFinSf2;
run;

/*scatter plots */
proc sgscatter data=trainclean;
matrix lgrlivarea llotarea l1stflrsf lmasvnrarea l2ndflrsf/group=scndflr;
run;

/*forward selection */
proc glmselect data =trainclean seed=49412531;
title 'Forward Selection Method';
class _character_;
model lsaleprice = lgrlivarea llotarea l1stflrsf bedroomabvgr mosold overallcond overallqual yearbuilt yearremodadd yrsold
_character_ /
selection=forward(stop=cv) cvmethod=random(15);
run;

proc glm data=trainclean plots=diagnostics;
class _character_;
title 'Model from Forward Selection';
model lsaleprice= lgrlivarea llotarea l1stflrsf BedroomAbvGr OverallCond OverallQual YearBuilt YearRemodAdd YrSold
MSZoning
Neighborhood BsmtQual BsmtExposure BsmtFinType1 CentralAir KitchenQual Functional GarageFinish SaleCondition
/solution clparm;
store forward;
run;

proc plm restore=forward;
show class;
score data = test2 out=predicteddata
pred=SalePrice;
run;

proc export data = predicteddata outfile='/home/u60173286/modelforward.csv' dbms=csv replace;
run;

/* backward selection */
proc glmselect data =trainclean seed=459568854;
title 'Backward Selection Method';
class _character_;
model lsaleprice = lgrlivarea llotarea l1stflrsf bedroomabvgr mosold overallcond overallqual yearbuilt yearremodadd yrsold
_character_ /
```

```
selection=backward(stop=cv) cvmethod=random(15);
run;

proc glm data =trainclean plots=diagnostics;
title 'Model from Backward Selection';
class _character_;
model lsaleprice=lgrlivarea llotarea l1stflrsf BedroomAbvGr MoSold OverallCond OverallQual YearBuilt YearRemodAdd YrSold
MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 BldgType MasVnrType
ExterQual
ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 Heating HeatingQC CentralAir KitchenQual
Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature
SaleCondition/solution clparm;
store backward;
run;

proc plm restore=backward;
show class;
score data = test2 out=predicteddata
pred=SalePrice;
run;

proc export data = predicteddata outfile='/home/u60173286/modelbackward.csv' dbms=csv replace;
run;

/* stepwise selection */
proc glmselect data =trainclean seed=290152413;
title 'Stepwise Selection Method';
class _character_;
model lsaleprice = lgrlivarea llotarea l1stflrsf bedroomabvgr mosold overallcond overallqual yearbuilt yearremodadd yrsold
_character_ /
selection=stepwise(select=cv) cvmethod=random(15);
run;

proc glm data =trainclean plots=diagnostics;
title 'Model from Stepwise Selection';
class _character_;
model      lsaleprice= lgrlivarea llotarea l1stflrsf BedroomAbvGr OverallCond OverallQual YearBuilt YearRemodAdd MSZoning
LotConfig LandSlope Neighborhood Condition1 MasVnrType BsmtQual BsmtExposure BsmtFinType1 Heating HeatingQC
CentralAir
KitchenQual Functional GarageType GarageFinish GarageCond PoolQC SaleCondition/solution clparm;
store stepwise;
run;

proc plm restore=stepwise;
show class;
score data = test2 out=predicteddata
pred=SalePrice;
run;

proc export data = predicteddata outfile='/home/u60173286/modelstepwise.csv' dbms=csv replace;
run;

/* custom model with interactions .13267 kaggle score (25.5%)*/
proc glmselect data=trainclean;
class _character_;
model lsaleprice = lgrlivarea llotarea l1stflrsf BedroomAbvGr MoSold OverallCond OverallQual
YearBuilt YearRemodAdd YrSold MSZoning Street Alley LandContour Utilities LotConfig
LandSlope Neighborhood Condition1 BldgType ExterQual ExterCond Foundation BsmtQual
BsmtCond BsmtExposure BsmtFinType1 Heating HeatingQC CentralAir KitchenQual Functional
GarageFinish GarageQual GarageCond PavedDrive PoolQC SaleCondition
age scndflr age*kitchenqual scndflr*l1stflrsf llotarea*l1stflrsf l2ndflrsf*scndflr scndflr*bedroomabvgr
vnr*lmasvnrarea bsmt bsmt*totalbsmtsf bsmt*bsmtfinsf2/ selection=backward(stop = cv include=70);
run;

proc glm data=trainclean plots=diagnostics;
class _character_;
model lsaleprice = lgrlivarea llotarea l1stflrsf BedroomAbvGr MoSold OverallCond OverallQual YearBuilt YearRemodAdd YrSold
MSZoning Street Alley LandContour Utilities LotConfig LandSlope Neighborhood Condition1 BldgType ExterQual ExterCond
Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 Heating HeatingQC CentralAir KitchenQual Functional
```

```
GarageFinish GarageQual GarageCond PavedDrive PoolQC SaleCondition age scndflr age*kitchenqual scndflr*l1stflrsf
llotarea*l1stflrsf l2ndflrsf*scndflr scndflr*bedroomabvgr vnr*lmasvnrarea bsmt bsmt*totalbsmtsf bsmt*bsmtfinsf2/ solution;
store custom;
run;

proc plm restore=custom;
show class;
score data = test2 out=predicteddata
pred=SalePrice;
run;
proc export data =predicteddata outfile='/home/u60173286/modelcustom11.csv' dbms=csv replace;
run;
```

Appendix 9 SAS Code for Analysis 2