Kaggle Project

# Introduction

“Ask a home buyer to describe their dream house, and they probably won’t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition’s dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.” ([Kaggle](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview))

## Data Description

The [Ames Housing dataset](http://jse.amstat.org/v19n3/decock.pdf) was compiled by Dean De Cock and describes residential property sales in Ames, Iowa from 2006 to 2010. “The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values.”[[1]](#footnote-1) [Full data description](http://jse.amstat.org/v19n3/decock/DataDocumentation.txt)

# Question 1

## Restatement of Problem

Century 21 Ames, a real estate company in Ames, Iowa, has commissioned us to estimate how the sales price of a house in the North Ames (NAmes), Edwards, and Brookside (BrkSide) neighborhoods is related to the square footage (GrLivArea) of the house and if the sales price and its relationship to square footage depends on which neighborhood the house is located in.

## Build and Fit the Model

In assessing the data, the original training dataset has been filtered down to only the neighborhoods of interest. It appears that a linear relationship exists between square footage and sale price, although a handful of outliers are also present. However, the log-log transformed data appeared to be a more appropriate fit for the data, so we proceeded using that transformation. **Figure 2**

Based on the problem, our model for the relationship between square footage and sale price is:

Equation 1: Model

Log Sale Price = βo + β1\*Log Living Area + β2\*Edwards + β3\*NAmes + β4\*Log Living Area∗Edwards + β5\*Log Living Area∗NAmes

A screenshot of a cell phone

Description automatically generated

Figure 1: Model Summary

The results of the linear model indicate that our final model for the North Ames, Edwards, and Brookside neighborhoods is as follows:

Equation 2: Final model for North Ames, Edwards & Brookside neighborhoods

Log Sale Price = 6.0068 + 0.8066\*Log Living Area + 2.0262\*Edwards + 2.4065\*NAmes − 0.2907\*Log Living Area∗Edwards − 0.3226\*Log Living Area∗NAmes

This model has an R-squared value of 0.5177 and an Adjusted R-squared value of 0.5113, thus 51.77% of the variation in sales price can be explained by the variation in neighborhood and square footage. These values are simply estimates; the 95% confidence intervals can be found in the appendix. (Figure 3)

### North Ames (NAmes)

Log Sale Price = 6.0068 + 0.8066\*Log Living Area + 2.4065 − 0.3226\*Log Living Area

Log Sale Price = 8.4133 + 0.484\*Log Living Area

### Edwards

Log Sale Price = 6.0068 + 0.8066\*Log Living Area + 2.0261 − 0.2907\*Log Living Area

Log Sale Price = 8.0341 + 0.5159 \*Log Living Area

### Brookside (BrkSide)

Log Sale Price = 6.0068 + 0.8066\*Log Living Area

## Checking Assumptions

### Examining Residuals (Figure 4)

1. **Linearity**: The plots of residuals and studentized residuals, although primarily clustered around 11.5-12.0, do appear as a random cloud.
2. **Normality**: The histogram shows the residuals to be fairly normally distributed.
3. **Equal Standard Deviation**: The Q-Q plot shows the residuals to be linearly distributed.
4. **Outliers**: There are a handful of outliers that we decided to include in this model.
5. **Cook’s D**: Based on the Cook’s D plot, most of the data is relatively clumped together, however there is a single outlier that we should be careful of. (Figure 5)

## Conclusion

Using the full model ANOVA (Figure 6) and the reduced model ANOVA (Figure 7), we can build our own ANOVA table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | Df | Sum sq. | Mean sq. | F value | Pr(>F) |
| Model | 2 | 0.6395096 | 3.12739 | 84.5924 | < .00001 |
| Error | 377 | 13.93775037 | 0.03697016 |  |  |
| Total | 379 | 14.57725997 |  |  |  |

Evidence suggests that there is a statistically significant difference between the slopes between neighborhoods controlling for living space (p-value < .00001). Selling a home increase the median sale price by 124% (p-value 2e-16), Edwards 67% (p-value .0015), and North Ames 62% (5.5e-5) for every 1 unit increase in median Living area respectively (p-value < 2.2e-16, Adjusted R-squared: 0.5113, & RMSE: 0.1896609).

# Question 2

## Restatement of the Problem

Build a predictive model for sales prices of all residential property sales in Ames, Iowa using multiple linear regression to analyze all of the variables in the dataset. In addition to building a custom model, we will also include three additional models: forward selection, backwards elimination, and stepwise selection. Finally, we will compare the Adjusted R-squared, CV Press, and Kaggle scores for each of the models to determine which offers the most accurate prediction of future home sales in Ames, Iowa.

## Models Considered

Four different models were constructed:

* Forward selection – modeled in SAS using proc glmselect
* Backward selection – modeled in SAS using proc glmselect
* Stepwise selection – modeled in SAS using proc glmselect
* Custom – modeled in SAS using proc glmselect

## Examining Residuals (Figure 11)

1. **Linearity**: The plots of residuals and studentized residuals, do appear as a random cloud.
2. **Normality**: The histogram shows the residuals to be fairly normally distributed.
3. **Equal Standard Deviation**: The Q-Q plot shows the residuals to be linearly distributed.
4. **Outliers**: There are a handful of outliers that we decided to include in this model.
5. **Cook’s D**: Based on the Cook’s D plot, most of the data is relatively clumped together, however there are a couple of outliers that we should be careful of.

## Model Selection

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Adjusted R2 | CV Press | Kaggle Score |
| Forward (Figure 8) | 0.8370 | 33.3403 |  |
| Backward (Figure 9) | 0.9390 | 35.9403 |  |
| Stepwise (Figure 10) | 0.8370 | 33.9818 |  |
| Custom | 0.9029 |  |  |

## Conclusion

# Appendix

## Question 1

A close up of a map

Description automatically generated

Figure 2: Square footage and sales price, original and log-log transformed

A screenshot of a cell phone

Description automatically generated

Figure 3: 95% confidence intervals for the model

A screenshot of a social media post

Description automatically generated

Figure 4: Log-log model residuals

A screenshot of a cell phone

Description automatically generated

Figure 5: Cook’s D analysis of original and log-log transformed data

A screenshot of a cell phone

Description automatically generated

Figure 6: Full model ANOVA

A screenshot of a cell phone

Description automatically generated

Figure 7: Reduced model ANOVA

### Question 1 Code

*# Setup--------------------------------------------------------------------*

*# Load libraries*

**library**(tidyverse)

**library**(hrbrthemes) *# clean plotting theme*

**library**(scales) *# format scales*

**library**(patchwork) *# organizing plots*

**library**(broom) *# for working with model*

**library**(bookdown) *# for working with captions & references*

**library**(knitr) *# for table formatting*

**library**(xtable) *# for exporting model results*

*# Load data*

test <- read\_csv("test.csv")

train <- read\_csv("train.csv")

*# Clean up column names*

test <- janitor::clean\_names(test)

train <- janitor::clean\_names((train))

*# Q1: Initial Scatter Plots------------------------------------------------*

*# Original data scatter plot*

a <- train %>%

filter(neighborhood %**in**% c("NAmes", "Edwards", "BrkSide")) %>%

ggplot(aes(gr\_liv\_area, sale\_price, color = neighborhood)) +

geom\_point(alpha = 0.3) +

geom\_smooth(method = "lm") +

scale\_y\_continuous(label = dollar) +

labs(title = "Kaggle Home Prices",

subtitle = "original data",

x = "Living Area Square Footage",

y = "Sale Price",

color = "Neighborhood") +

theme\_ipsum() +

theme(legend.position = "bottom") +

NULL

*# Log-log data scatter plot*

b <- train %>%

filter(neighborhood %**in**% c("NAmes", "Edwards", "BrkSide")) %>%

ggplot(aes(log(gr\_liv\_area), Log Sale Price , color = neighborhood)) +

geom\_point(alpha = 0.3) +

geom\_smooth(method = "lm") +

scale\_y\_continuous(label = dollar) +

labs(title = "Kaggle Home Prices",

subtitle = "with log-log transformation",

x = "log Living Area Square Footage",

y = "log Sale Price",

color = "Neighborhood") +

theme\_ipsum() +

theme(legend.position = "bottom") +

NULL

*# Patchwork layout*

a + b

*# Q1: Model Construction---------------------------------------------------*

*# Filter training set*

train\_by\_neighborhood <- train %>%

filter(neighborhood %**in**% c("NAmes", "Edwards", "BrkSide")) %>%

mutate(sq\_foot = round(gr\_liv\_area, digits = -2),

log\_sq\_foot = log(sq\_foot),

log\_sale\_price = Log Sale Price ) %>%

select(id, sq\_foot, log\_sq\_foot, neighborhood, sale\_price, log\_sale\_price)

*# Identify outliers*

outliers <- boxplot(train\_by\_neighborhood$log\_sale\_price, plot = FALSE)$out

train\_by\_neighborhood\_outliers <- train\_by\_neighborhood %>%

filter(!log\_sale\_price %**in**% outliers)

*# Create models*

log\_model <- lm(log\_sale\_price ~

log\_sq\_foot +

neighborhood +

log\_sq\_foot\*neighborhood,

data = train\_by\_neighborhood)

log\_model\_reduced <- lm(log\_sale\_price ~

log\_sq\_foot +

neighborhood)

*# Full model statistics*

summary(log\_model) *# summary statistics*

confint(log\_model) *# confidence intervals*

anova(log\_model) *# ANOVA*

*# Reduced model statistics*

anova(log\_model\_reduced) *# ANOVA*

*# Q1: Plot Residuals-------------------------------------------------------*

*# Plot residuals*

a <- log\_model %>%

ggplot(aes(.fitted, .resid)) +

geom\_point(shape = 1, alpha = 0.5) +

geom\_hline(yintercept = 0, color = "blue") +

labs(title = "Residuals") +

theme\_ipsum()

*# Plot studentized residuals*

b <- log\_model %>%

augment() %>%

ggplot(aes(.fitted, .std.resid)) +

geom\_point(shape = 1, alpha = 0.5) +

geom\_hline(yintercept = 0, color = "blue") +

labs(title = "Studentized Residuals") +

theme\_ipsum()

*# Q-Q Plot of Residuals*

c <- log\_model %>%

ggplot(aes(sample = .resid)) +

stat\_qq(alpha = 0.5) +

stat\_qq\_line(color = "darkblue") +

labs(title = "Q-Q Plot of Residuals") +

theme\_ipsum()

*# Histogram of residuals*

d <- log\_model %>%

ggplot(aes(.resid, ..density..)) +

geom\_histogram(fill = "lightblue", color = "darkblue") +

geom\_density() +

labs(title = "Histogram of Residuals") +

theme\_ipsum()

*# Assemble plots using patchwork*

(a + b) / (c + d)

*# Q1: Cook's D-------------------------------------------------------------*

log\_model %>%

augment() %>%

ggplot(aes(.fitted, .cooksd)) +

geom\_jitter(alpha = 0.3) +

labs(title = "Cook's D of Log Model") +

theme\_ipsum()

## Question 2

A screenshot of a cell phone

Description automatically generated

Figure 8: Forward selection model results

A screenshot of a cell phone

Description automatically generated

Figure 9: Backward selection model results

A screenshot of a cell phone

Description automatically generated

Figure 10: Stepwise selection model results

A close up of a map

Description automatically generated

Figure 11: Fit diagnostics

A screenshot of a cell phone

Description automatically generated

Figure 12: SAS residuals

A screenshot of a cell phone

Description automatically generated

Figure 13: SAS residuals

### Question 2 Code

**proc** **import** datafile = '/folders/myfolders/SASInput.csv'

out = work.Train

dbms = CSV replace

;

**run**;

**proc** **import** datafile = '/folders/myfolders/SASOutput.csv'

out = work.Test

dbms = CSV

;

**run**;

**DATA** TEST\_Log;

SET TEST;

LogSalePrice = **.**;

**run**;

**DATA** TRAIN;

SET TRAIN (DROP = GRLIVAREA);

LIVING = ROUND(LIVING,**100**);

**RUN**;

**PROC** **SQL**;CREATE TABLE NEIGHBORHOODS AS SELECT DISTINCT NEIGHBORHOOD FROM TRAIN;**QUIT**;

**DATA** AMES\_IOWA;

SET TRAIN

(WHERE=(INDEX(NEIGHBORHOOD,'AMES') > **0** OR INDEX(NEIGHBORHOOD,'IOWA') > **0**));

**RUN**;

ods graphics on;

/\*

Forward

\*/

**proc** **glmselect** data=Train TESTDATA= work.TEST\_Log plots=all;

CLASS Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2

BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual

Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond

GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual

LandContour LandSlope LotConfig LotFrontage LotShape MSZoning MasVnrType

MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition

SaleType Street Utilities

;

model LogSalePrice = MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape

LandContour 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 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF

Heating HeatingQC CentralAir Electrical \_1stFlrSF \_2ndFlrSF LowQualFinSF

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars

GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch

ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition

Living LogLiving

/ selection = Forward(select = ADJRSQ stop = CV SLE = **.15**) /\*details=all stats=all\*/;

**run**;

/\*

Backward

\*/

**proc** **glmselect** data=Train TESTDATA= work.TEST\_Log plots=all;

CLASS Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2

BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual

Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond

GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual

LandContour LandSlope LotConfig LotFrontage LotShape MSZoning MasVnrType

MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition

SaleType Street Utilities

;

model LogSalePrice = MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape

LandContour 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 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF

Heating HeatingQC CentralAir Electrical \_1stFlrSF \_2ndFlrSF LowQualFinSF

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars

GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch

ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition

Living LogLiving

/ selection = Backward(select = ADJRSQ stop = CV CHOOSE = CV) /\*details=all stats=all\*/;

**run**;

/\*

Stepwise

\*/

**proc** **glmselect** data=Train TESTDATA= work.TEST\_Log plots=all;

CLASS Alley BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2

BsmtQual CentralAir Condition1 Condition2 Electrical ExterCond ExterQual

Exterior1st Exterior2nd Fence FireplaceQu Foundation Functional GarageCond

GarageFinish GarageQual GarageType Heating HeatingQC HouseStyle KitchenQual

LandContour LandSlope LotConfig LotFrontage LotShape MSZoning MasVnrType

MiscFeature Neighborhood PavedDrive PoolQC RoofMatl RoofStyle SaleCondition

SaleType Street Utilities

;

model LogSalePrice = MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape

LandContour 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 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF

Heating HeatingQC CentralAir Electrical \_1stFlrSF \_2ndFlrSF LowQualFinSF

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars

GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch

ScreenPorch PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition

Living LogLiving

/ selection = STEPWISE(select = ADJRSQ stop = CV slentry = **.15** SLE = **.15**) /\*details=all stats=all\*/;

**run**;

/\*

Custom Model

\*/

**proc** **glm** data = Train PLOTS=all;

CLASS Neighborhood MSZoning HouseStyle CentralAir GarageFinish SaleCondition

BsmtQual ExterQual KitchenQual LotFrontage BsmtFinType1;

model LogSalePrice = LogLiving Neighborhood MSZoning

HouseStyle GarageArea GarageCars OverallQual

TotalBsmtSF CentralAir GarageFinish ExterQual \_1stFlrSF FullBath YearBuilt YearRemodAdd SaleCondition KitchenQual

BsmtQual BsmtFinType1 BsmtFinType1\*TotalBsmtSF TotRmsAbvGrd LotFrontage Neighborhood\*LogLiving/ solution;

**run**;

ods graphics off;

1. Dean De Cock, “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project”, Journal of Statistics Education, Volume 19, Number 3(2011). [↑](#footnote-ref-1)