Ames Iowa Housing price prediction

SMU MSDS Statistical Foundations Project



Joey Hernandez, Derek Rogers, Nicole Assenza

TAble of  
Contents

Exploratory Data Analysis 1

Introduction 1

Transformations 1

Analysis 1 3

Restatement of Problem 3

Build and Fit the model 3

Checking Assumptions 4

Influential point analysis 5

R Shiny: Price vs. Living area 6

Conclusion 7

Analysis 2 8

Restatement of problem 8

Model Selection 8

Forwards Selection 9

Backwards Selection 9

Stepwise Selection 10

Custom Model 11

Comparing Competing Models 12

Conclusion 12

Appendix 13

# Exploratory Data Analysis

## Introduction

When purchasing a home many variables play a role in determining the value and price an individual is willing to pay. Our First objective is to assist Century 21 - Ames in understanding how the sale price of a home is related to the living area. Additionally, we would like to investigate and understand if the sale price and its square footage relationship is dependent on the neighborhood it’s located in.

Secondly, we will construct a predictive model for the sales prices of homes in all of Ames, Iowa. To ensure the best model implementation we will create 4 models and evaluate them on their R^2, CV PRESS, and Kaggle Score statistics.

## Transformations

The dataset was compiled by Dean De Cock for use in data science education and describes almost every aspect of residential homes in Ames, Iowa. The “Train” dataset contains 1460 rows and 81 Columns. To find out more on this data and problem set [click here](https://www.kaggle.com/c/house-prices-advanced-regression-techniques).

To ensure that we are working with data that will yield accurate and actionable results, we engaged in visualization and testing of variables to determine if they met the assumptions of regression or if they required intervention. The assumptions of linearity, homoscedasticity, independence, and normality were assessed for each of the variables that were measured in square footage and dollars. Below is a table that summarizes our findings on the variables that required a log transformation or failed to meet the assumptions of linear regression. The R code used to gather the evidence for these conclusions can be found [Here](https://joeyhdz.github.io/projects/Cyclone_Final/).

Table

Description automatically generatedTable

Description automatically generated

Figure 1

Figure 2

# Analysis 1

## Restatement of Problem

Century 21 Ames sells homes in the North Ames, Edwards, and Brookside neighborhood and would like to investigate how the sale price of the house is related to the square footage of the living area of the house and if the sale price (and its relationship to square footage) depends on which neighborhood the house is located in.

## Build and Fit the model

The initial model we used to visualize the relationship between Sale price and Square footage is:

Est. SalePrice = ßo + ß1 GrLivArea

the following are the next two models we chose after further analysis:

Est. Median Log(SalePrice) = ßo + ß1Log(GrLivArea)

Est. Median Log(SalePrice) = ßo + ß1log(GrLivArea)+ß2 Neighborhood + ß3 Neigh \* GrLivArea

## Checking Assumptions

Chart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

Figure 3

Figure 4

The “Residual vs Predicted Values” and “Residual Histogram” for the log of sales price has a few extreme residuals but follows a normal distribution.

Linearity: A linear relationship does seem appropriate to describing the data, although due to some outliers, this does not seem to hold constant above 3000 ft^2.

Normality: Normality seems to be intact. There is some evidence of outliers in the dataset, but given the sample size we will cite the Central Limit theorem and proceed to investigating all other assumptions

Constant Variance: There is a cluster of residuals and they do not seem to be equally spread.

Independence: We will assume independence, but will note that the homes we are investigating are within a specific set of neighborhoods and certain confounding variables could present themselves such as same company of construction workers with certain home buy/sell relationship, and neighborhoods with certain levels of wealth will attract various buyers/sells that can influence the pricing for one neighborhood to another etc.

Given the evidence of non-constant variance and present outliers, we will proceed with a LOG-LOG transformation which addresses our concerns and will allow us to present accurate and interpretable findings.

## Influential point analysis

Chart

Description automatically generatedTimeline

Description automatically generated with medium confidence

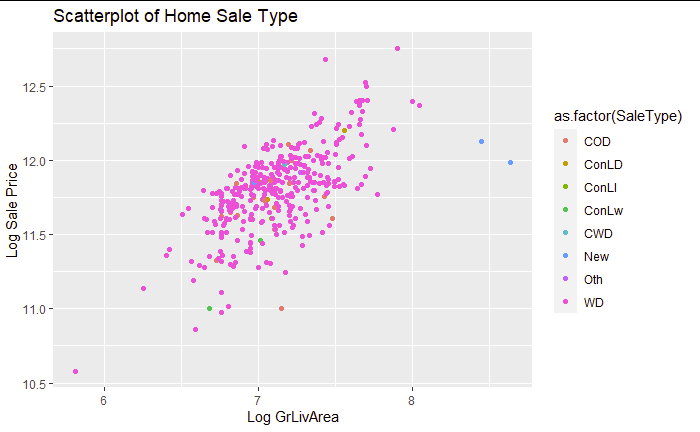
Figure 6

Figure 5

In plotting GrLivArea v SalePrice we found that there are some influential points in the data.

The Cook's D chart shows a point that is an extreme outlier. So that we can better understand this observation we will investigate the outlier to see if it makes sense to keep or remove it.

Chart, scatter chart

Description automatically generated

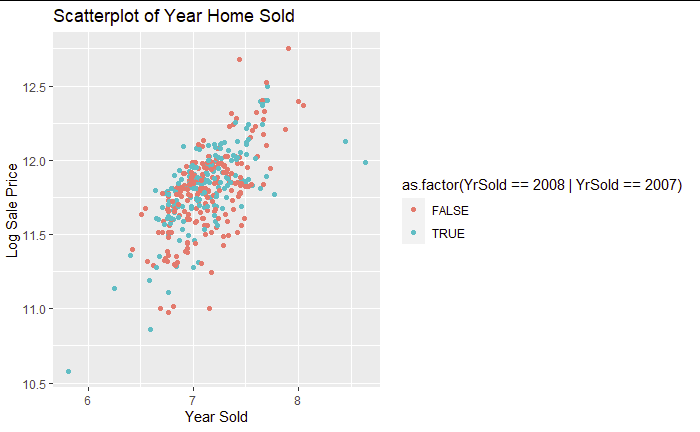


Figure 9

Figure 8

Figure 7

upon further investigation of the 2 outliers that we are assessing. It appears that both homes were of a "New" sale type, during 2007 and 2008, both of which are years in which new homes would be expected to have sold for losses or abnormal pricing due to the recession and housing market collapse. The two homes were also of a sale condition "Parital". a "Partial" Sale condition only makes up 1% of the home sale condition types within the home sales for these 3 neighborhoods.

In conclusion, the abnormality of the sales of these two homes seems to be attributed to circumstances that are not what is to be expected under normal market conditions and predictions. In addition to the justification to leave out our 2 discovered outliers, the data without the outliers also better fits the model assumptions for linear regression. We will proceed without these two points.

Next, we will build a model for QOI 2 that will determine if the square footage of the living area and its Sale Price is dependent on which Neighborhood it is in. To do this it will help first to test if all the regression lines are parallel. if they are, then we can check for a difference in Sale Price across all Neighborhoods by looking at the intercept.

Our analysis of the full and reduced model uses the adjusted R^2 result which favors the model with the interaction included, thus, we will proceed with the full model.

**Table

Description automatically generated**

Figure 10

## R Shiny: Price vs. Living area

An app visualizing Sale Price v Living Area can be found [**here**](https://joeyhdz.shinyapps.io/Ames-Housing-App/?_ga=2.185711593.1513787878.1670416282-1021813642.1667347631)

**Chart, scatter chart

Description automatically generated**

Figure 11

## Conclusion

The log of the sales price has a plausible linear relationship with the log of the living area with strong evidence for the assumptions of linear regression. Explanations for influential points in the dataset were analyzed, understood, and removed from the model. Finally, through evidence it was determined that we should proceed with the full model giving us a conclusive linear regression equation.

# Analysis 2

## Restatement of problem

Removing the narrow focus of the previous analysis, we move onto building a model that predicts the sales price for any house in Ames Iowa using all of the data. We will start with a linear regression model that utilizes all of the variables we deemed fit for model building. Next, various auto selection algorithms are run to provide a useful subset of variables. Finally we will analyze the results from the selection algorithms and build a custom model that maximizes predictability, confidence in the model, and interpretability.

## Model Selection

The selected features for our core evaluation were determined by correlative relationships, as well as features which were intuitive to what a home buyer may look for. After compiling our core feature set, we then processed those features through various selection methods listed below with the goal of identifying the simplest model that performs well on the data. This approach can help reduce the complexity of the model and improve its interpretability, as well as reduce the risk of overfitting.

## Forwards Selection

Forward selection begins with no variables selected (the null model). In the first step, it adds the most significant variable. At each subsequent step, it adds the most significant variable of those not in the model, until there are no variables that meet the criterion set by the user.

Table

Description automatically generated

Figure 12

## Backwards Selection

Backward selection begins with all the variables selected, and removes the least significant one at each step, until none meet the criterion.

Table

Description automatically generated

Figure 13

## Stepwise Selection

Stepwise selection alternates between forward and backward, bringing in and removing variables that meet the criteria for entry or removal, until a stable set of variables is attained.

Table

Description automatically generated

Figure 14

## Custom Model

Table

Description automatically generated

Figure 15

## Comparing Competing Models

Table

Description automatically generated

Figure 16

## Conclusion

We created 4 different models using linear regression to predict sale prices of houses in Ames, and we have found that our custom model performed the best. We took nine predictor variables out of the 81 (OverallQual, log(GrLivArea), Neighborhood, GarageCars, ExterQual, TotalBsmtSF, GarageArea, KitchenQual and yearBuilt) and developed those into a regression model with surprisingly good results.  We ended up with an R^2 value of .88, meaning that 88% of the variance of the sales price variable is explained by the variation in our nine predictors variables.

# Appendix

[**For a detailed look at our code and visuals please click here**](https://joeyhdz.github.io/projects/Cyclone_Final/)

Graphical user interface, text, application

Description automatically generated