​

​ ​

Introduction

This case study will focus on the analysis of the price of homes, and on creating predictive models for the final sale price of homes in Ames, Iowa.

The first analysis question asks us to look into the relationship between the sales price of homes and the square footage of those homes per three neighborhoods of interest.

As for the second analysis question, we will use regression modeling and other techniques in order to select the best explanatory variables for the optimal model that will predict any home’s sale price.

Data Description

The open-source data were provided by Kaggle’s site for the House Prices - Advanced Regression Techniques ***https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data****.* The data set for the training data is made up of 1460 observations, while the test data set is comprised of 1459 observations with 79 explanatory variables. Since the data are readily available on Kaggle’s website, one can find the data set description, training, and testing data sets under the tab labeled Data.

Analysis Question 1

Restatement of Problem

As part of our 1st analysis, we will be examining how the sale price of homes is related to the square footage of those homes with respect to 3 neighborhoods of interest. Those neighborhoods happen to be Edwards, North Ames (NAmes), and Brookside (BrkSide).

Preparation

We start by reading in the dataset and looking at the structure of both the training and testing datasets. Then, we round the living area (GrLivArea) by increments of 100 sq. ft. and visualize the data to get an initial understanding of the relationship between SalePrice and RoundedLivArea. Also, we filter the data to include only the three neighborhoods of interest: NAmes, Edwards and BrkSide.

Build and Fit the Model

We built a linear regression model using the original training dataset and checked for assumption violations. We found that the residuals were not normally distributed with an unequal variance so we performed a log transformation on both SalePrice and RoundedLivArea.

After the transformation, the assumptions seem to be met. However, there were two outliers that needed to be removed to improve the model.

Chart, scatter chart

Description automatically generated

Fig.1: Scatterplots depict original data without transformation of Sale Price vs. Rounded Living Area per Neighborhood

Chart, scatter chart

Description automatically generated

Fig. 2: Scatterplots depict log-log transformation of Sale Price vs. Rounded Living Area per Neighborhood

Transformation

By having performed a log transformation of the data from Fig.1, we will address some concerns of unequal variance that appear to be present in Fig.1. As we see from Fig.1, from each data set based on neighborhood, the overall spread for each neighborhood is shown to start off large and tapper off to smaller variances. These types of violations will be addressed through a log-log transformation. Upon having implemented the log transformation that we can see in Fig.2, we are able to perceive visually strong evidence of linearity and improved constant variance. In other words, we can see a relationship between the log of Sale Price and the log of Rounded Living Area with respect to the Neighborhoods (BkrSide, Edwards, NAmes) of interest.

Nota Bene

It is critical to mention that the data pre- and post-log transformation do contain some observations of interest. Those observations—or houses of interest—appear to be outliers that have high leverage with a possibility of high or low influence. We will address those concerns later on in our study.

Furthermore, we would like to address that though the log transformation does appear to create a slight skewness in the data, the concern is appeased due to our sampling size (n) which is indeed > 30 and thus we would rely on Central Limit Theorem.

Check Assumptions

In reviewing the residual plots for the log transformed data—see Fig.2—the relationship between the predictor variable and the target variable appears to be linear. Each observation appears to be independent. The variance of each observation are consistent across all the values of the predictor variables. The observations appear to be normally distributed.

Influential point analysis (Cook’s D and Leverage)

There are two outliers that are uniquely set apart from the rest of the data; their relationships between Sale Price and Living Area (pre- and post-log transformation) do not behave likewise to the rest of the observations shown within Fig.1 and Fig.2. Upon a closer examination of those two potential outliers, it was found that both of the outliers had a high living area in terms of sq.ft. (>3800 sq.ft. in the original scale or about 8.25 on the log scale).

Comparing Competing Models

We will start by building various models using different combinations of predictors and interactions. Then, we shall calculate R2 and Adjusted R2 for each model, which provides insights into how well the model fit, while penalizing it for complexity. The higher the R2, the better the fit, taking into account the number of predictors in the model from the Adjusted R2. The R2 and adjustedR2 are higher in the Model with interactions which can be seen in Fig.3 and Fig.4.

To improve our model and avoid overfitting, we will utilize Internal Cross Validation PRESS. The Internal CV PRESS value was 14.22, indicating a reasonably good predictive performance when compared to alternative models. The value suggests that our model can effectively generalize to new observations while minimizing the risk of overfitting.



Fig.3a : PRESS Value

**Without Interactions**

Table

Description automatically generated

Fig.3b Model without Interactions

**With Interactions**

Table

Description automatically generated

Fig.4 Model with Interactions

Referencing at Fig.3b and Fig.4 above, we will have to compare both models, the one without interactions and the one with interactions, to find out what the coefficients are and how significant each is for their respective model, with a significance level of 0.05.

There is sufficient evidence to suggest, with a significance level of 0.05, that the full model is significantly better than the reduced model (p-value  = 0.0002749). We would have to reject our null hypothesis in favor of our alternative hypothesis, and therefore conduct our study using the full model.

Text

Description automatically generated

Fig.5 Full Model and Reduced Model

Text

Description automatically generated

Fig.6: Lack of Fit Test (Full Model and Reduced Model)

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated with low confidence

Fig.7: Residuals Analysis

Assumptions

Linearity: Upon examining the scatter plot, we can see that there appears to exist a strong linear trend between Sale Price and Living Area. The mean of those distributions appears to follow a linear pattern.

Normality: By examining the QQ Plot, we can see that overall—given that the sample size (n) is greater than 30—normal distribution albeit a slight skewness.

Equal Standard Deviation: There looks to be significant evidence for homoscedasticity.

Independence: As far as the information provided, and the fact that there are house Identifications to go along with Sale Prices, independence will be assumed for each observation.

Nota Bene

It is important to note that, from the scatter plot, the Residuals v. Fitted shows signs of a randomized cloud; this event favors the usage of linear modeling.

Parameters

Chart, scatter chart

Description automatically generated

Fig.8 Regression models with Intervals of Log Sale Prive v. Log Rounded Living Area by Neighborhood

Table

Description automatically generated

Text

Description automatically generated

Fig. 9: This captures the estimates, standard errors, t-values, and p-values

 of our variables used in our model. This also includes our Confidence Intervals.

Interpretation

**Intercept Interpretation:**

We estimate that the median sale price for a house with a living area of 0 sq.ft. in the Brookside neighborhood to be associated with 27.454=  $175.34 (p-value < 2e-16). We are 95% confident that the sales price is between 26.98787039 = $126.93 and 27.92093889 = $242.35.

**Slope Interpretation:**

We estimate that for the Edwards neighborhood, there is an 2-0.01601 = 0.99 times multiplicative decrease in the median sale price for a house (p-value = 0.616). This is with respect to the Brookside neighborhood. We are 95% confident that the multiplicative decrease is within the interval of 0.95 = 2-0.07865432 and 1.03 = 20.04663353 or an 5% decrease to 3% increase in comparison to Brookside.

We estimate that for the North Ames neighborhood, there is an 2e0.12543 = 1.09 times multiplicative increase in the median sale price for a house (p-value = 1.37e-05). This is with respect to the Brookside neighborhood. We are 95% confident that the multiplicative increase is within the interval of  1.05=  20.06946555 and 1.13 = 2 0.18139252 or between 5% and 13% increase in comparison to Brookside.

Lastly, for Brookside we estimate that there is an 20.60079 = 1.52 times multiplicative increase in the median sale price for a house (p-value < 2e-16) in Brookside. We are 95% confident that the true multiplicative increase is between 20.53483012 = 1.45 and 20.66675200 =1.60 or between 45% and 60% increase.

Conclusion

In conclusion, our study found strong statistical evidence that the neighborhood where a home is located not only affects the sale price of the home, but also influences the relationship between living area and sale price.

Analysis Question 2

Restatement of Problem

For the second analysis, we aim to build the most predictive model for house sale prices in Ames, Iowa, considering all neighborhoods. The objective is to create four different models using various feature selection methods: forward selection, backward elimination, stepwise selection, and a custom model. The custom model can be any of the previous three models or a new one built by adding or subtracting variables as desired. The models will be evaluated using adjusted R2, CV Press, and Kaggle Score to determine which model performs best in predicting future sale prices of homes in Ames, Iowa.

Preparation

Cleaning up the data is foremost and the most crucial step of our study for analysis two. Upon reviewing the data, the data underwent a thorough process to check for NAs and addressed other values, such as missing variables.

Model Selection

**Backward Elimination**: The automation of the algorithm allows us to start with all candidate variables. It eliminates each variable one at a time until the AIC reaches its threshold where then it cannot remove another variable.

**Forward Selection:** This process involves starting with no variables in the model and then, through testing, adds another variable so long as it meets a criterion. This process is repeated until AIC cannot be lowered through the addition of a new variable.

**Stepwise**: In itself, this process is a combination of the two selection processes mentioned above. It initiates by adding the first most significant variable and then removes or ads one variable at a time till the AIC criterion is met.

**Custom**:

Type of Selection

* Custom Model

Checking Assumptions

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

A picture containing chart

Description automatically generated

Fig.11 The above 5 images are

* Influential Point Analysis (Cooks D and Leverage)

**Linearity**: Upon close examination of the residual plots, it is possible to see that the data are strongly penchant for a linear relationship. As there is a randomized cloud, the relationship between sale price in dollars and living area in terms of sq.ft. is reflected within the data.

**Normality**: This foremost assumption—quite pivotal, alongside linearity, is that upon examining the QQ plot for residuals, there are favorable indications for normality. The sample size is quite large, greater than 30, and we can say that the Central Limit Theorem would address any slight concern against normality.

**Equal Standard Deviation:** Upon examination of the residual plots, there seems to be a penchant for homoscedasticity.

**Independence:** As far as the information provided, and the fact that there are house Identifications to go along with Sale Prices, independence will be assumed for each observation.

Comparing Competing Models

* Adj R2
* Internal CV Press
* Kaggle Score

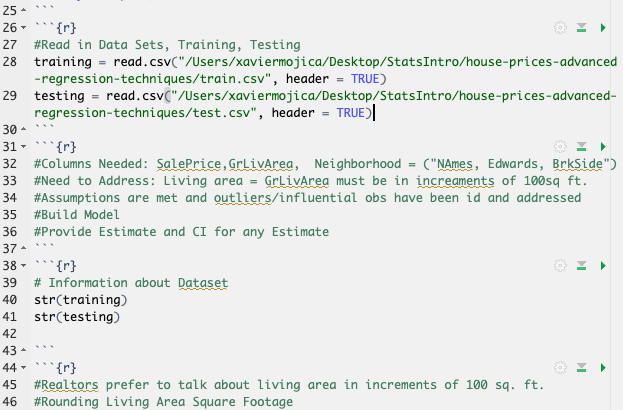
Conclusion

A short summary of the analysis.

Appendix A: Analysis 1

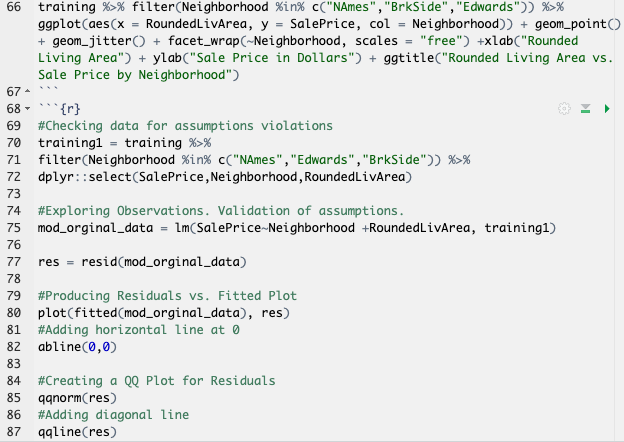
Graphical user interface, text, application, chat or text message

Description automatically generated



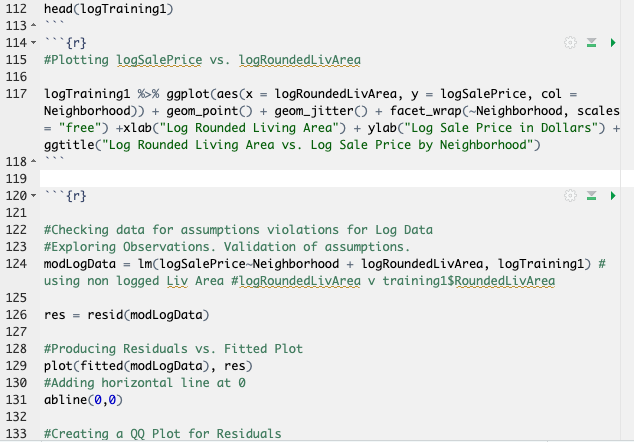
Graphical user interface, text, application, email

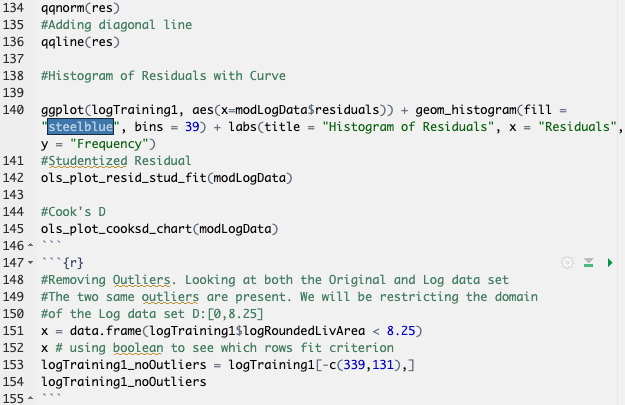
Description automatically generated



Graphical user interface, text, application, email

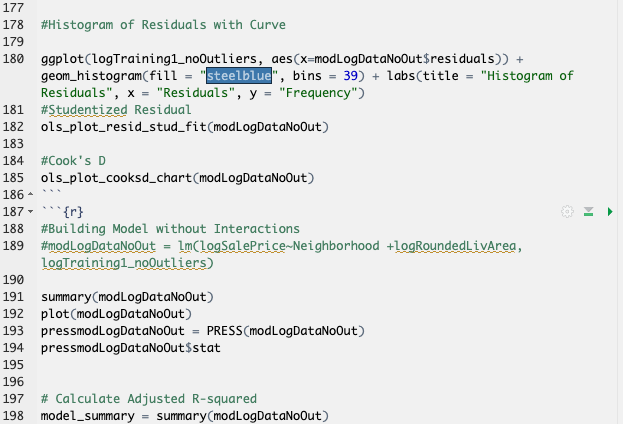
Description automatically generated

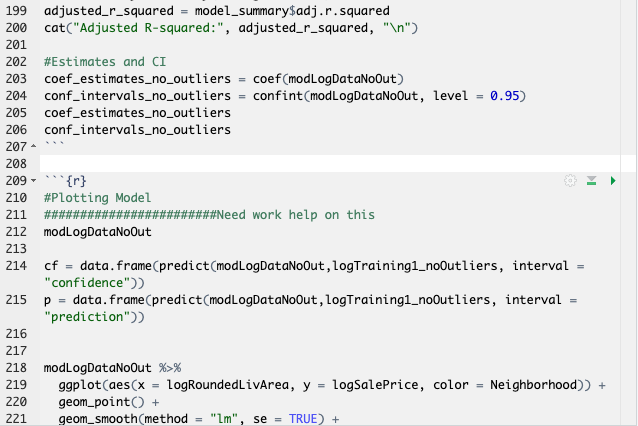




Graphical user interface, text, application, email

Description automatically generated





Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Appendix B: Analysis 2