Team Members: Andy Ho, An Nguyen, Matthew Norton

MSDS6371 – Kaggle Project – Assignment 13 and 14

Date: 8/12/18

# Introduction

## Data Descriptive Statistics

* File Name
  + test.csv (461 KB)
  + train.csv (461 KB)
* File Origin
  + <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>
* Data Dimensions

|  |  |  |
| --- | --- | --- |
|  | **Observations** | **Variables** |
| **Train** | 1460 | 81 |
| **Test** | 1459 | 80 |

* Selected Variables
  + Analysis 1
    - **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
    - **GrLivArea**: Above grade (ground) living area square feet
  + Analysis 2
    - <need to input>

# Analysis 1

## Introduction

Century 21 Ames only sells home in NAmes, Edwards, and BrkSide neighborhoods and would like to get an estimate of how the SalePrice of each house is related to the square footage of the living area of the house (GrLivArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located. This portion of the report will build a model with the provided dataset and provide our conclusion that quantifies the relationship between living area and sale price with respect to the three neighborhoods.

* Averages by neighborhoods and all three combined

|  |  |  |  |
| --- | --- | --- | --- |
| **Neighborhood** | **Number of Properties** | **Avg. Living Area** | **Avg. Sale Price** |
| NAmes | 225 | 1310.31 | $145,847.08 |
| Edwards | 100 | 1340.04 | $128,219.70 |
| BrkSide | 58 | 1203.07 | $124,834.05 |
| All Three | 383 | 1301.83 | $138,062.50 |

## Summary

Using the 383 observations for the three neighborhoods, our goal is to select a model that best fits the relationship between living area and sale price. This section will provide details on the selection process, the assumptions, and the model.

## Selection Process

This section will demonstrate the various models ran, results of each, and our decision on which model is best based on how it satisfies the following assumptions:

1. Normality
2. Linearity
3. Equal Variance
4. Independence
5. Outliers

## Summary Table of Model Assumptions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Vanilla Regression** | | | | | | |
| **Model** | **Normality** | **Linearity** | **Equal Variance** | **Independence** | **Outliers** | **R^2** |
| Original Data | ✔ |  |  | ✔ |  | 0.447 |
| Log-Linear | ✔ |  | ✔ | ✔ |  | 0.465 |
| Linear-Log | ✔ |  |  | ✔ |  | 0.466 |
| Log-Log | ✔ | ✔ | ✔ | ✔ |  | 0.512 |
| Original Data (Outliers Addressed) | ✔ | ✔ | ✔ | ✔ | ✔ | 0.523 |
| Log-Linear (Outliers Addressed) | ✔ | ✔ | ✔ | ✔ | ✔ | 0.525 |
| Linear-Log (Outliers Addressed) | ✔ |  | ✔ | ✔ | ✔ | 0.0498 |
| Log-Log (Outliers Addressed) | ✔ |  | ✔ | ✔ | ✔ | 0.053 |

## Assumptions

After running 8 different models, there is sufficient evidence that the vanilla regression model using the original data (three neighborhoods) while addressing the outliers is the best fit model. The following addresses the assumptions. **Refer to Addendum 5 for more charts and values**.

1. **Normality:** Judging by the histogram, the original data (outliers addressed) looks slightly better than the rest and shows evidence of normality.
2. **Linearity:** Judging by the scatterplots, the original data (outliers addressed) looks slightly better than the other two outliers addressed (log-linear and linear-log). There is sufficient evidence for linearity.
3. **Equal Variance:** Judging by the residual scatterplots and QQplot, the original data (outliers addressed) looks better than the rest and shows sufficient evidence for equal variance.
4. **Independence:** We will assume independence, although the data gathering process was not explained.
5. **Outliers:** There are a few outliers that were addressed. Specifically, homes in the Edwards neighborhood with living area greater than 4,000 sq. ft and two homes (one in Edwards and NAmes) with sale price over $300,000.

## Residual Plots



## Influential Point Analysis (Cook’s D and Leverage)



## Model

The model assigns SalePrice as the response variable and GrLivArea as the explanatory variable. Based on the scatterplots, there is strong evidence the model will have different slopes and intercepts for each significant parameter.

**Predicted SalePrice = ß0 + ß1 (GrLivArea) + ß2 (GrLivArea \* NAmes) +**

**ß3 (GrLivArea \* Edwards) + ß4 (GrLivArea \* BrkSide)**

## Fitted Model

## Model Metrics

### Adj R^2

### Internal CV Press

## Parameters

### Estimates

### Interpretation

### Confidence Intervals

## Conclusion

After using the vanilla regression and addressing the outliers of a living area greater than 4,000 sq. ft. and a sales price of over $300,000, there is significant evidence that both the neighborhood and square footage of the living area influence the estimated sale price of the home in Ames. The intercept in this model provides an estimated sale price of $80,327 for a house with a living area of one hundred square feet in the NAmes neighborhood. For a house with a living area of one hundred square feet, the model gives us an estimated sales price decrease of $60,354 for the BrkSide neighborhood and an estimated sales price decrease of $43,225 for the Edwards neighborhood. For every increased living area of one hundred square feet the estimated sales price increases $4,956 for a home in the NAmes neighborhood, the estimated sales price increases $3,780 in the BrkSide neighborhood and the estimated sale price increases $2,059 in the Edwards neighborhood.

PredictedSalesPrice = ß0 + ß1 \* (BrkSide) + ß2 \* (Edwards) + ß3 \* (GrLivAreaper100) + ß4 \* (BrkSide \* GrLivAreaper100) + ß5 \* (Edwards \* GrLivAreaper100)

PredictedSalesPrice = $80,325 – $60,354 \* (BrkSide) – $43,225 \* (Edwards) + $4,956 \* (GrLivAreaper100) + $3,760 \* (BrkSide \* GrLivAreaper100) + $2,059 \* (Edwards \* GrLivAreaper100)

NAmes Neighborhood:

PredictedSalesPrice= $80,325 + $4,956 \* GrLivAreaper100

BrkSide Neighborhood:

PredictedSalesPrice= $19,971 + $8,716 \* GrLivAreaper100

Edwards Neighborhood:

PredictedSalesPrice= $37,100 + $7,015 \* GrLivAreaper100

------------------------------------------------------------------------------------------------------------------------------------------

Predicted SalePrice = ß0 + ß1 (GrLivAreaper100) + ß2 (GrLivAreaper100\* NAmes) + ß3 (GrLivAreaper100\* Edwards) + ß4 (GrLivAreaper100\* BrkSide)

Predicted SalePrice = $80,325 + $4,956 \* (GrLivAreaper100) + $0 \* (GrLivAreaper100\* NAmes) + $2,059 \* (GrLivAreaper100\* Edwards) + $3,760 \* (GrLivAreaper100\* BrkSide)