Raag Patel & Ranjan Karki

Professor Dr. Bivin Sadler

MSDS 6371

August 1, 2021

**MSDS 6371 Final Project**

The dataset given to us, is compiled of a little less than 1500 home sales in Ames, Iowa. The data set contains 81 variables, detailing important details about the houses sold in that area, including the value of the sale, the amenities, features, and physical properties of the house. In the first analysis question, we utilized the price of the sale, the neighborhood, and the square footage of the living area, to fully answer the question given to us from Century 21 Ames.

Our model building for the entry into the Kaggle competition started by utilizing all 80 variables (not including the ID). We selected using statistical measures certain columns that included categorical and continuous data to build a model that is indicative of the price of home sales in Ames, Iowa.

We highlight some column names throughout the report, such as GrLiveAreaPer100, which means the square footage of the living area, in increments of 100.

**Analysis Question 1:**

Century 21 Ames, a real estate company in Ames Iowa, has commissioned us to help them answer some questions they had about prices in certain neighborhoods. We were asked to use a dataset that contained about 1500 home sales and narrow it down to certain neighborhoods. Where we then should build a model that highlights the relation between sale price of a home and the square footage of the living area of the house, depending on the neighborhood the house is located in. We were asked to provide confidence intervals in our analysis, and a thorough statistical analysis of the question of interest.

We ran 4 different models which are linear model, log-linear, linear log, model with outliers addressed. The model with outliers addressed is the best fitting model.

Model with outliers addressed:

PredictedSalesPrice = ß0 + ß1 \* (GrLivAreaPer100) + ß2 \* (Edwards) + ß3 \* (North Ames) + ß4 \* (Edwards \* GrLivAreaPer100) + ß5 \* (North Ames \* GrLivAreaPer100)

Normality: Looking at the histogram of residuals there is no evidence that the residuals do not follow normal distribution. Given the size of the sample, it is enough to follow the central limit theorem.

Linearity: Log transformed, and untransformed model residual plot looks random cloud providing enough evidence that linearity assumptions are met. It is obvious that transformation didn’t help. So proceeded with untransformed data with outliers addressed. This evidence is backed by QQ plot which shows most of the data are linearly related.

Equal Variance: From the residual’s plots most of the data points are scattered above and below the reference line indicating equal variance. However, there is clustering effect seen. It is most likely most of the houses are in the same price ranges. Also, from QQ plot the data points are normally distributed in relation with x-axis.

Independence: We will assume the data points are independent for the purpose of the question. House prices are often related to each other, as certain areas are deemed more “wealthy”. Since we cannot measure the impact of that, we can and will assume independence.

Outliers: There were few outliers that were addressed by using Cooks D-method. Any data points that were above 3 times the mean was removed. There were only 9 observations as outlier, when we removed these outliers adjusted R2 improved significantly.

**Before outliers addressed After outliers addressed**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

The influential points were analyzed using Cooks D. Any data points that were 3 times the mean was as considered outliers and new model was fitted after removing these data points. This significantly improved Residual vs Leverage plot. After removing there were no data points that was high leverage low studentized residual. Please refer to the plot above.

Normality: Judging from the histogram of the residuals, with the outliers removed, there is no evidence that residual do not follow normal distribution.

Linearity: The residual plot looks randomly distributed and there is sufficient evidence for linearity.

Equal variance: QQ plot - that data points are normally distributed in relation with x-axis, sufficient evidence of equal variance.

Independence: We will assume the data is independent to each other, by ignoring the clustering effect.

|  |  |
| --- | --- |
| **Model** | **Adj R2** |
| Original | 0.44 |
| Log-Log | 0.5056 |
| Linear-Log | 0.4587 |
| Outliers Addressed | 0.5226 |

Setting Brookside as reference;

PredictedSalesPrice = ß0 + ß1 \* (GrLivAreaPer100) + ß2 \* (Edwards) + ß3 \* (North Ames) + ß4 \* (Edwards \* GrLivAreaPer100) + ß5 \* (North Ames \* GrLivAreaPer100)

Fitted Model;

PredictedSalesPrice = 22396.3 + 8416.6\* (GrLivAreaPer100) + 35252.2\* (Edwards) + 55475.3 \* (North Ames) - -3235.5 \* (Edwards \* GrLivAreaPer100) -3247.7 \* (North Ames \* GrLivAreaPer100)

A screenshot of a computer

Description automatically generated with medium confidenceGraphical user interface, text

Description automatically generated

Three regression equations:

Predicted (SalesPrice|Brookside) = 22396.3 + 8416.6 \* (GrLivAreaPer100)

Predicted (SalesPrice|Edwards) = 57648.5 + 5181.1 \* (GrLivAreaPer100)

Predicted (SalesPrice|North Ames) = 77871.6 + 5168.9\* (GrLivAreaPer100)

To answer the question about how the square footage of the living area relates to the sale price of the home, dependent on neighborhood, we can refer to the regression equations we provided earlier.

For homes in Brookside, we estimate that for every 100 feet of square footage added, the price of houses sold increased by $8,416.60, from a base of $22,396.30.

For homes in Edwards, we estimate that for every 100 feet of square footage added, the price of houses sold increased by $5,181.10 from a base of $57,648.50.

For homes sold in North Ames, we estimate that for every 100 feet of square footage added, the price of houses sold increased by $5,168.90 from a base of $77,871.60.

From the above analysis, as client Century 21 targets the average sales price in relation to living area for the three neighborhoods. Removing outliers helped us to narrow down the average living space with average sales price, because there were the houses with more than 4000 square feet for less than $170,000 which is possibility of foreclosure. Whereas some houses with smaller than 1700 square foot living area were more than $300,000.

**R Shiny: Price v. Living Area Chart**

[Link to RShiny App.](https://raagpatel1.shinyapps.io/RShiny/) The app is useful to show scatterplots of the sale price of the home vs the square footage of the home, including the outliers we had removed to build the model. There are 3 plots for the 3 neighborhoods Century 21 Ames asked us to look at regarding their model. The sale price of the homes is not adjusted, as the model utilized the standard data as well.

**Analysis Question 2**

* Restatement of Problem

We were also asked to create the most predictive model we can for the sales price of Ames Iowa.

* Model Selection
* Type of Selection
* Stepwise
* Forward
* Backward
* Custom
* Checking Assumptions
* Residual Plots
* Influential point analysis (Cook’s D and Leverage)
* Make sure to address each assumption
* Comparing Competing Models