**Data Description**

(Where did the data come from? How big is it? How many observations? Where can

we find out more? What are the specific variables that we need to know with respect to

your analysis?)

**Analysis Question 1:**

* Restatement of Problem

Century 21 Ames would like to find how the living area( GrLIvArea) square footage is related to the SalesPrice in the Names ,Edwards and BrkSide neighborhoods. In this Boston Housing data set ,there are 383 observations for these three neighborhoods. We need to find the best model that explains relationship between living area and sale price.

* Build and Fit the Model

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 \* (NAmes) + ß4 \* (Edwards \* GrLivAreaper100) + ß5 \* (NAmes \* GrLivAreaper100)

* Checking Assumptions

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 to each other.

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.

Residual Plots

**Before outliers addressed After outliers addressed**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

* Influential point analysis (Cook’s D and Leverage)

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.

* Make sure to address each assumption.

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.

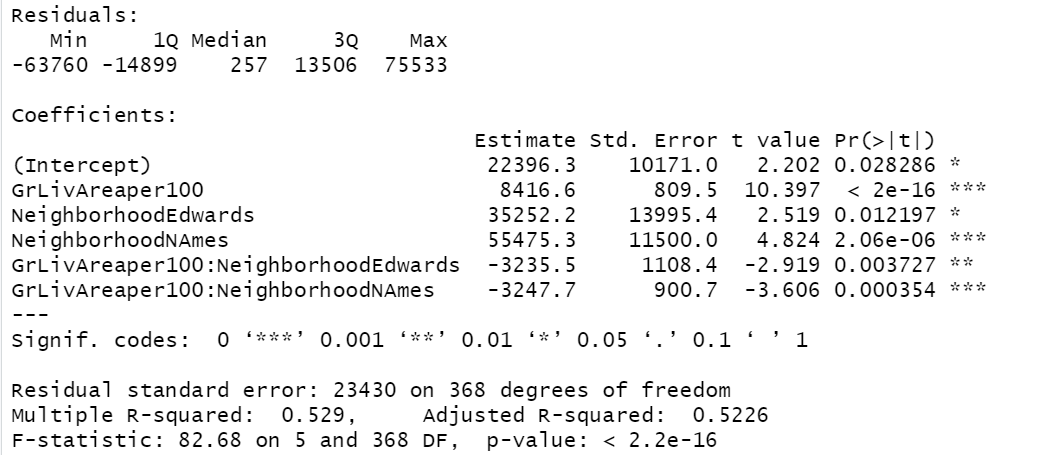
Equal variance: QQ plot - that data points are normally distributed in relation with x-axis, sufficient evidence of equal variance.

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

* Comparing Competing Models

|  |  |
| --- | --- |
| Model | Adj R2 |
| Original | 0.44 |
| Log-Log | 0.5056 |
| Linear-log | 0.4587 |
| Outlier Addressed | 0.5226 |

* Parameters Estimates And Interpretation



**Setting Brkside as reference**

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

**Fitted Model**

PredictedSalesPrice = 22396.3 + 8416.6\* (GrLivAreaper100) + 35252.2\* (Edwards) + 55475.3 \* (NAmes) - -3235.5 \* (Edwards \* GrLivAreaper100) -3247.7 \* (NAmes \* GrLivAreaper100)

Three regression equation:

Predicted(SalesPrice|Brkside) = 22396.3 + 8416.6 \* (GrLivAreaper100)

Predicted(SalesPrice|Edwards) = 57648.5 + 5181.1 \* (GrLivAreaper100)

Predicted(SalesPrice|NAmes) = 77871.6 + 5168.9\* (GrLivAreaper100)

Interpretations

The estimated average sale price for houses in the Brkside neighborhood is $22396.35 with no living area considered. For every 100 unit in living area in Brkside neighborhood, the estimated sales price increases by $8416.6

The estimated average sale price for houses in the Edwards neighborhood is $57648.5 with no living area considered. For every 100 unit in living area in Edwards neighborhood, the estimated sales price increases by $5181.1

The estimated average sale price for houses in the NAmes neighborhood is $77871.6 with no living area considered. For every 100 unit in living area in NAmes neighborhood, the estimated sales price increases by $5168.9.

* Confidence Intervals / Conclusion

Graphical user interface, text

Description automatically generated

A short summary of the analysis.

The intercept of the model provides the estimated average sales price for Brkside neighborhood is $22396.3 when living area is not considered. And we are 95% confident that price is between ($2395.768, $42396.774) for BrkSide neighborhood. For every 100 unit in living area in Brkside neighborhood, the estimated sales price increases by $8416.6

For house without living area considered the average estimated sales price in Edwards neighborhood is increased by $57648.5 with 95 % confidence interval of ($ 7731.154,$62773.281) with respect to Brkside.For every 100 unit in living area in Edwards neighborhood, the estimated sales price decreases by $3235.5 than BrKSide.

For house without living area considered the average estimated sales price in NAmes neighborhood is increased by $54704.888 with 95 % confidence interval of ($32861.295,$78089.322) with respect to Brkside. For every 100 unit in living area in NAmes neighborhood, the estimated sales price decreases by $3247.7 than Brkside.

**Conclusion**: 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 foot 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**

**Analysis Question 2**

* Restatement of Problem
* 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