**MSDS6371 Project**

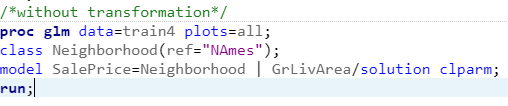
**(Kaggle)House Prices: Advanced Regression Techniques**

**Analysis 1:**

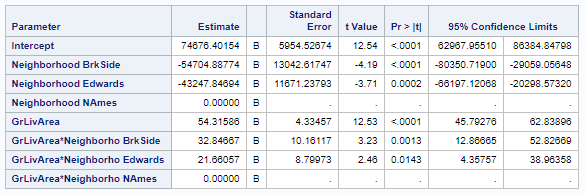
In the first analysis, the train data is downloaded from Kaggle competition. Our task is to predict the sale price of houses in three neighborhoods, with explanatory variable, categorical variable and dependent variable as GrLivArea, Neighborhood and SalePrice, respectively.

**Analysis the data quickly, identify outliers and remove outlier.** The cook’s D, residual, qq plot improve after removal of outliers, see appendix 1.

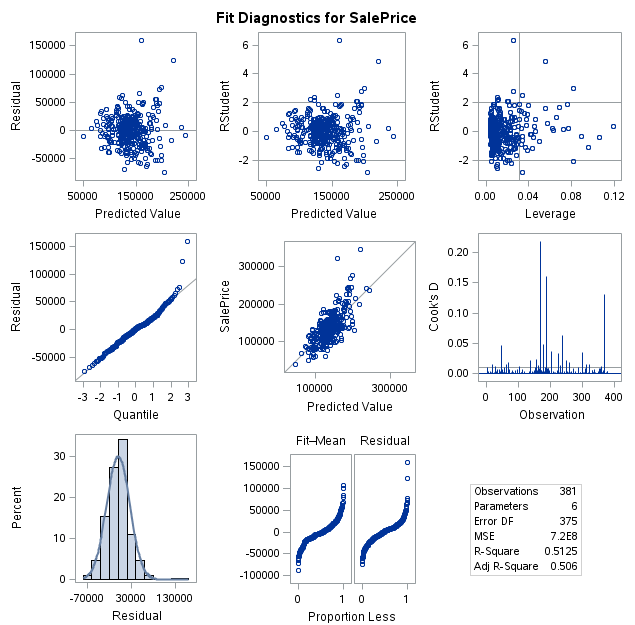
**Before data transformation:**



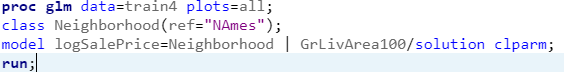
Model before data transformation:





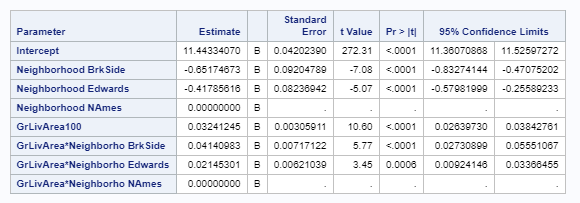


**After data transformation:**



Model after data transformation:

Parameters: Estimates and Confidence Intervals, interpretation in the conclusion.





**Overall Model:** log(SalePrice)=11.4433-0.6517\*BrkSide-0.4179\*Edwards+(GrLivArea/100)\*0.0324+ (GrLivArea/100)\*BrkSide\*0.0414+ (GrLivArea/100)\*Edwards\*0.0215

**In Neighborhood** **BrkSide:**

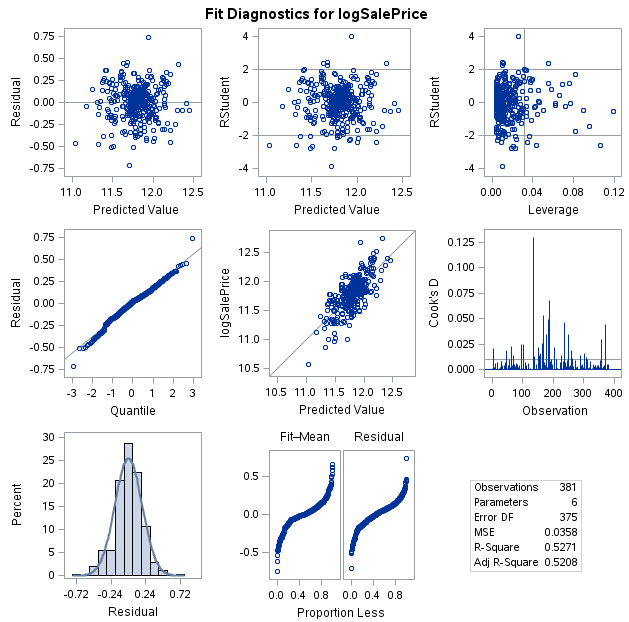
log(SalePrice)=11.4433-0.6517\*1 +(GrLivArea/100)\*0.0324+ (GrLivArea/100)\*1 \*0.0414= 10.7916+(GrLivArea/100)\*0.0738

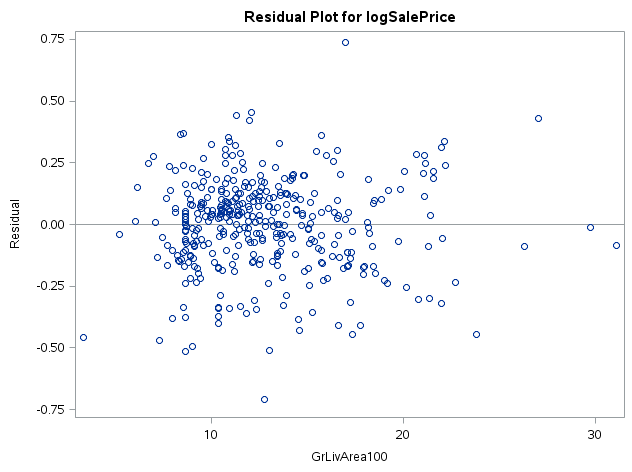
**In Neighborhood** **Edwards:**

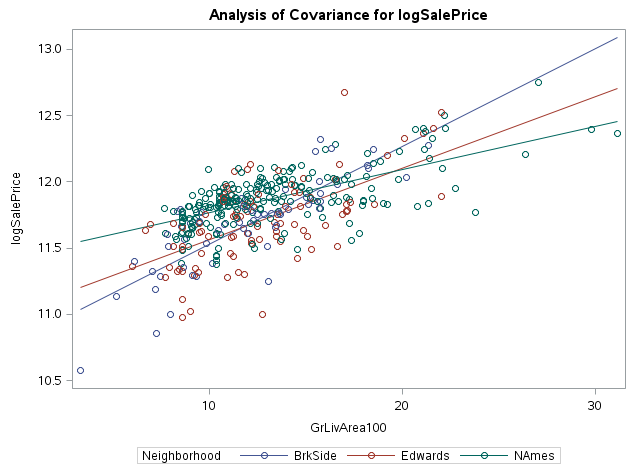
log(SalePrice)=11.4433-0.4179\*1+(GrLivArea/100)\*0.0324+ (GrLivArea/100)\*1\*0.0215 = 11.0254+(GrLivArea/100)\*0.0539

**In Neighborhood NAmes:**

log(SalePrice)=11.4433 +(GrLivArea/100)\*0.0324







**Check Assumptions:**

**Linearity:** From scatterplot, it shows linearity.

**Normality:** From the qqplot, no evidence against normality.

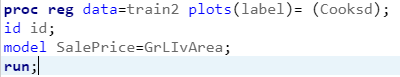
**Constant variance:** the residual plot is scattered, no obvious evidence against constant variance.

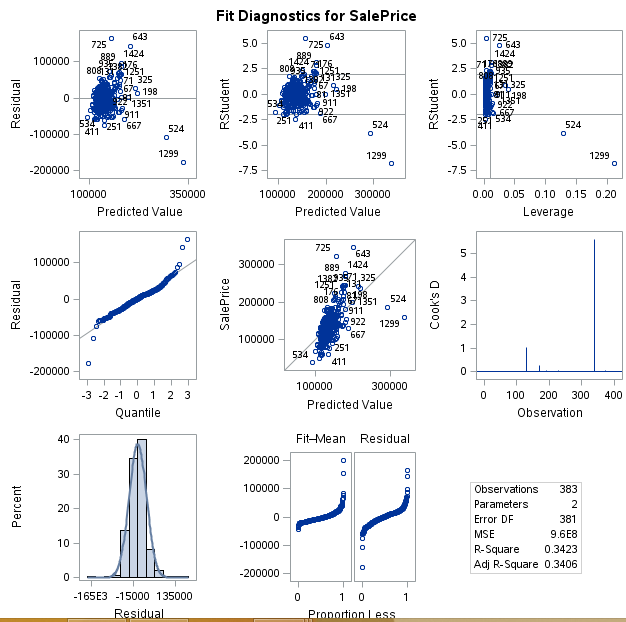
**Independence:** we can assume independency, and proceed with cautious.

**Conclusion**: The sale price of any house in these three neighborhood can be predicted by the models as above. Starting at sale price of , a 100 square foot increase of above grade (ground) living area is associated with a multiplicative change in median of sale Price in the Neighborhood of BrkSide. Starting at sale price of , a 100 square foot increase of above grade (ground) living area is associated with a multiplicative change in median of sale Price in the Neighborhood of Edwards. Starting at sale price of , a 100 square foot increase of above grade (ground) living area is associated with a multiplicative change in median of sale Price in the Neighborhood of NAmes.

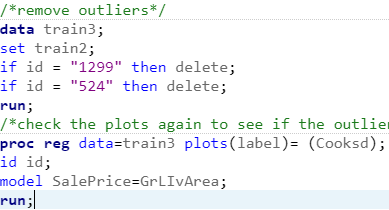
**Appendix 1:**

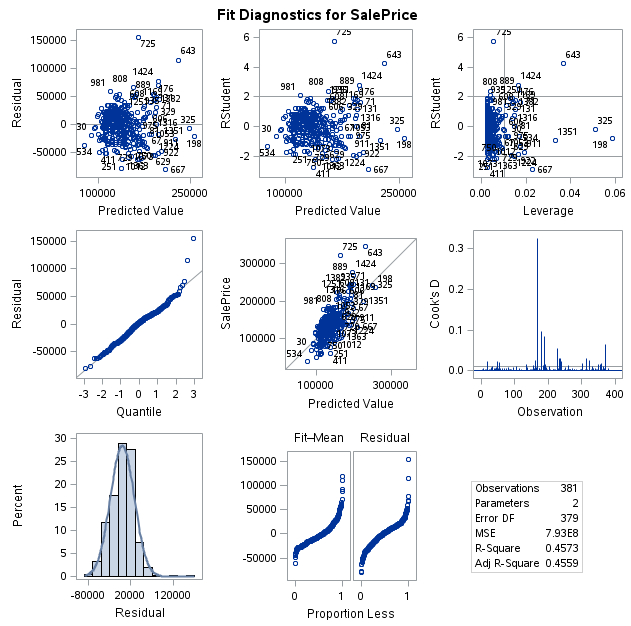
**Check outliers:**



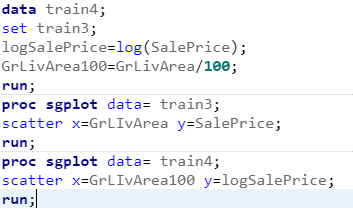


**After outliers removal,** the cook’s D plot and residual, qq plot improved.





**Appendix 2:**



|  |  |
| --- | --- |
| Before data transformation | After data transformation |
|  |  |

**Appendix 3.**

