**MSDS 6371 Project**

Renfang Wang 1; Venkata ‘MG’ Vanga1

# Introduction

The project work is part of academic program for Masters in Data Science (MSDS 6371). Data form Kaggle (<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>) is considered for the project and two analyses are carried out:

1. Analysis one – to analyze the relationship between sales price and living area for neighborhoods (‘*NAmes’, ‘Edwards’ and ‘Brkside’*).
2. Analysis two – to analyze the and predict the sales price of the given test data considering all the neighborhoods and other variables.

The results and conclusions are described in detail below. Both SAS and R are used to visualize the results.

# Description of Data

As indicated in the introduction the data is taken from Kaggle website. The data is for the Iowa state house sales prices and the attributes that effect the sales price in Iowa state. There are 80 variables in total with 1460 observations for each variable (few observations in some variables which do not have values are indicated as ‘NA’). Sales price, Ground Living Area, Neighborhood are few of main attributes in the data that are considered for the analysis.

# Analysis Question #1

## **Problem**

To find how sales price of a house is related to ground living area in the neighborhoods of ‘*NAmes’, ‘Edwards’ and ‘Brkside’*. The real estate company that is looking for this relationship is Century 21 Ames.

## Build and Fit the Model

The sales price V/s. ground living area initial scatter plot for the three *‘NAmes’, ‘Edwards’ and ‘Brkside’* neighborhoods is shown in Appendix - Figure 1. Regression lines for each neighborhood are indicated using distinct colors.

The regression line that fits all the three neighborhoods is indicated with **Black** line and the regression equation for the line is shown. The 95% confidence intervals and 95% prediction intervals for the regression is indicated using red and blue dashed lines respectively.

From initial inspection of Appendix - Figure 1, it is observed that:

1. There is a positive relationship between sale price and the ground living area for all three neighborhoods.
2. The intercepts and slopes of the regression lines for the three neighborhoods are different.
3. There are four significant outliers of which three are from *‘Edwards’* neighborhood and one is from *‘NAmes’* neighborhood.
4. The initial fit (black) regression line is **not a good fit** for the model and further evaluation is needed. Only 34.2% of the relationship between sales price and living area is explained by the regression line fit.

Since the intercepts and slopes are different, it is decided to use an interaction for the GrLivArea and Neighborhood.

The details of the four outliers (shown in Appendix - Figure 1) and the anticipated reason for outliers are summarized in the Appendix - Table 1. Based on the summary it is decided to treat Outlier 1 and 2 as a separate group and proceed with the analysis considering outlier 3 and 4 in the analysis.

## Assumption Checks

The Residual plots for the interaction model are shown in Appendix - Figure 2. Observations based on the residual plots are given below:

1. The Residuals v/s. Predicted Value plot shows few outliers, but since have already discussed about these residuals, we proceed cautiously proceed with the analysis. Also, there is an even scatter of the residuals with no curvature which suggest there is a linear trend and equal standard deviations.
2. There are few leverage points from the Rstudent v/s. Leverage plot.
3. Judging from the q-q plot and histogram of the residuals, there is no evidence that the residuals do not follow a normal distribution with constant variance.
4. Cook’s D also show 2 extreme values.
5. Though there are few leverage and influential points analysis we proceed with caution. We assume that the data are independent.

## Comparing Models

The R2, Adj. R2, CV PRESS and Predicted R2 values are shown in Appendix - Figure 3. The values indicate the interaction model is a good model for the analysis considering the three neighborhoods as categorical variables to predict the sales price depending on the ground living area.

## Model Parameters

The final scatter plot, regression equation and estimates are shown in Figure 1.

The regression equation for the interaction model:

Regression Equation for Neighborhood – ‘**NAmes**’:

The mean sales price in ‘NAmes’ neighborhood can be interpreted as, there is a predicted increase of $5,432 for every 100 sq. ft increase in Ground living area.

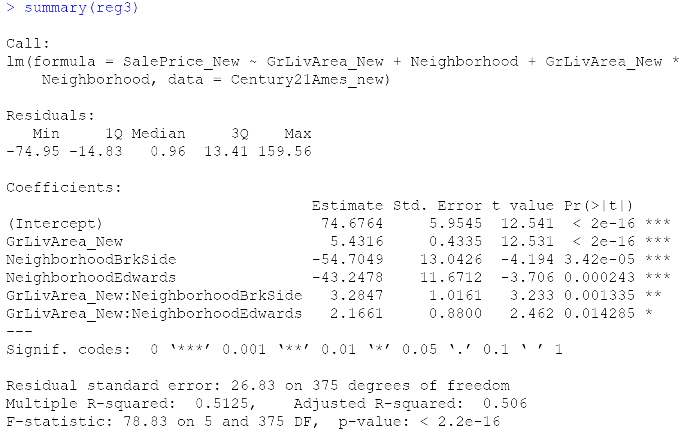
Regression Equation for Neighborhood – ‘**BrkSide**’:

The mean sales price in ‘BrkSide’ neighborhood can be interpreted as, there is a predicted increase of $8,176 for every 100 sq. ft increase in Ground living area.

Regression Equation for Neighborhood – ‘**Edwards**’:

The mean sales price in ‘Edwards’ neighborhood can be interpreted as, there is a predicted increase of $7,598 for every 100 sq. ft increase in Ground living area.

*The mean sales price is predicted to be approximately $172,000 when the Ground Living Area is greater than 35,000 sq. ft when the sale condition is partial (when sold before full construction is complete).*



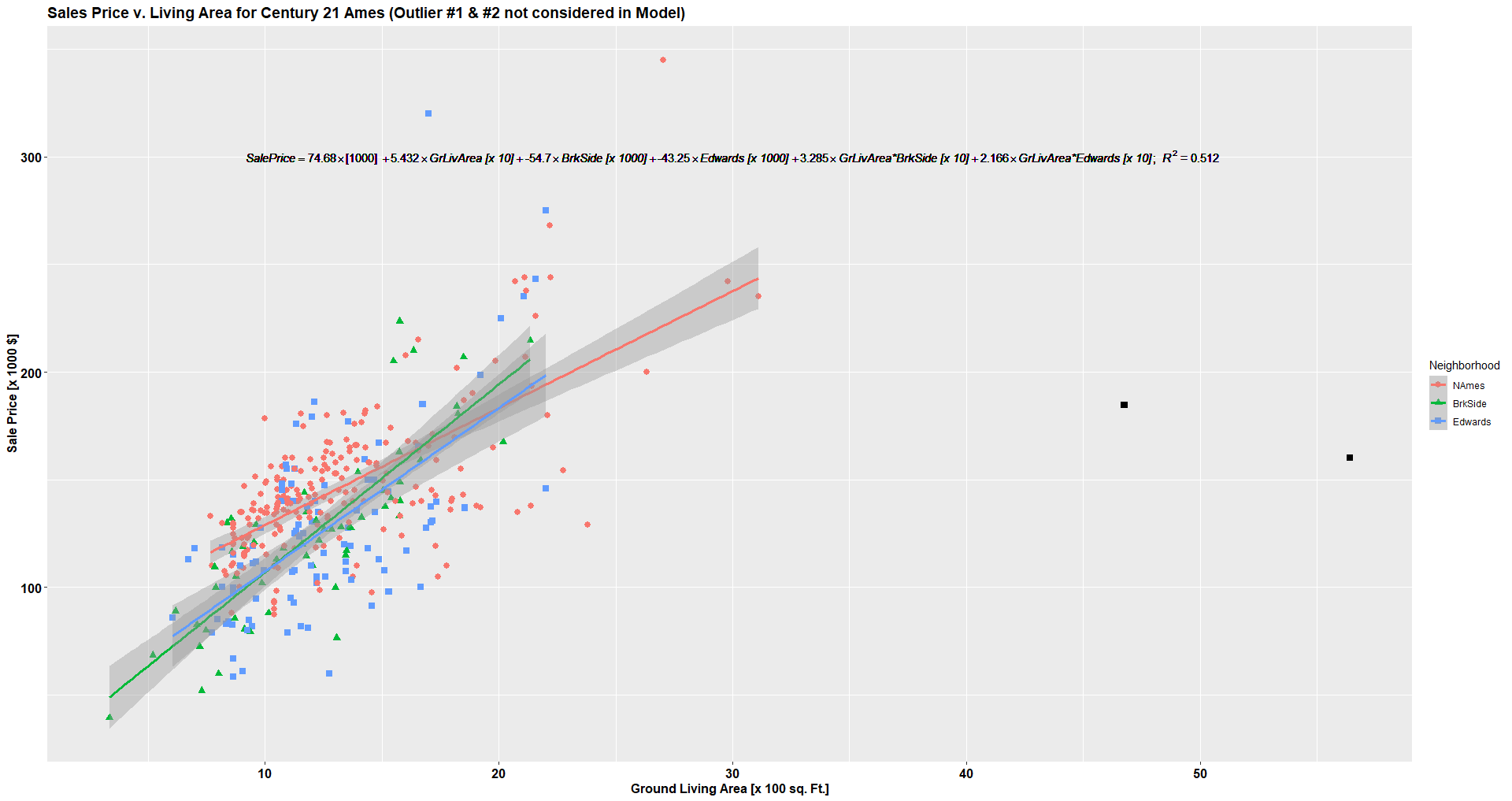
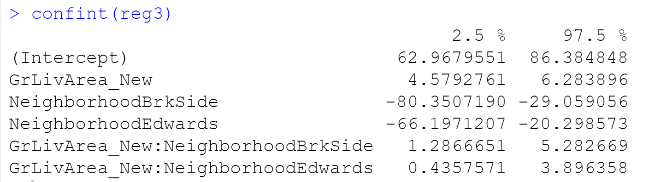


Figure 1: Model Scatter plot Sale Price v/s. Living Area, Regression Equation and Parameters (Analysis 1)

Evaluating the parameters shown in Figure 1, it can be observed that all the relevant variables and interactions used for the model are statistically significant (p-value < 0.05).

The 95% confidence intervals for the mean sale price increase are [$4,579 – $6,284] per 100 sq. ft in ‘NAmes’ neighborhood.

The 95% confidence intervals for the mean sale price increase are [$5,866 – $11,567] per 100 sq. ft in ‘BrkSide’ neighborhood.

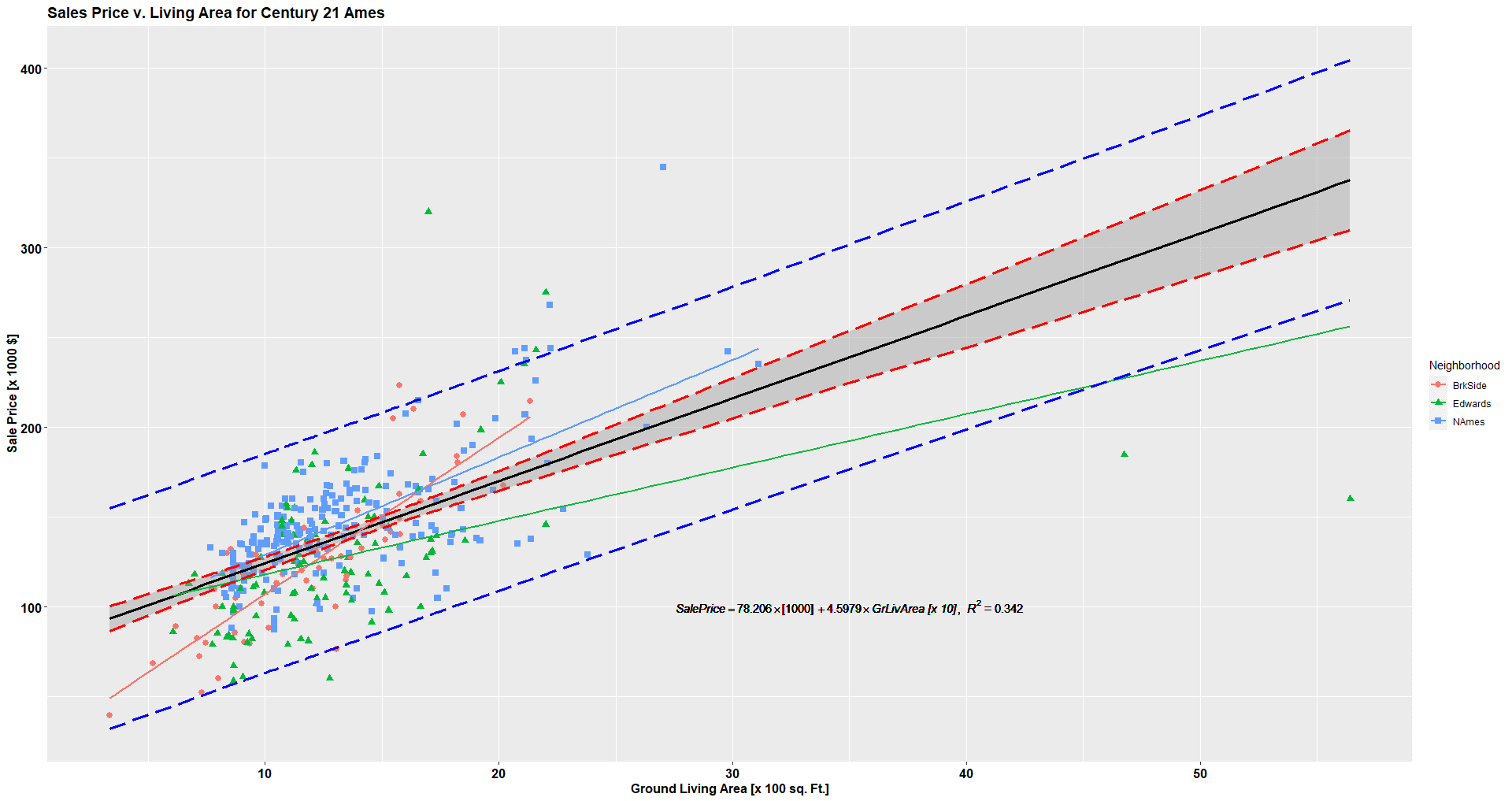
The 95% confidence intervals for the mean sale price increase are [$5,015 – $10,180] per 100 sq. ft in ‘Edwards’ neighborhood.

## **Conclusion**

The mean predicted sales prices are evaluated based on the ground living area for the neighborhoods of ‘NAmes’, ‘BrkSide’ and ‘Edwards’. The mean sale price evaluation equations and the 95% confidence intervals indicate that ‘NAmes’ neighborhood ($5,432) has the lowest increase in predicted mean sale price per every 100 sq. ft increase in ground living area followed by ‘Edwards’ neighborhood ($7,598) and ‘BrkSide’ neighborhood having the highest increase ($8,176).

# Analysis Question #2

Appendix A



1

2

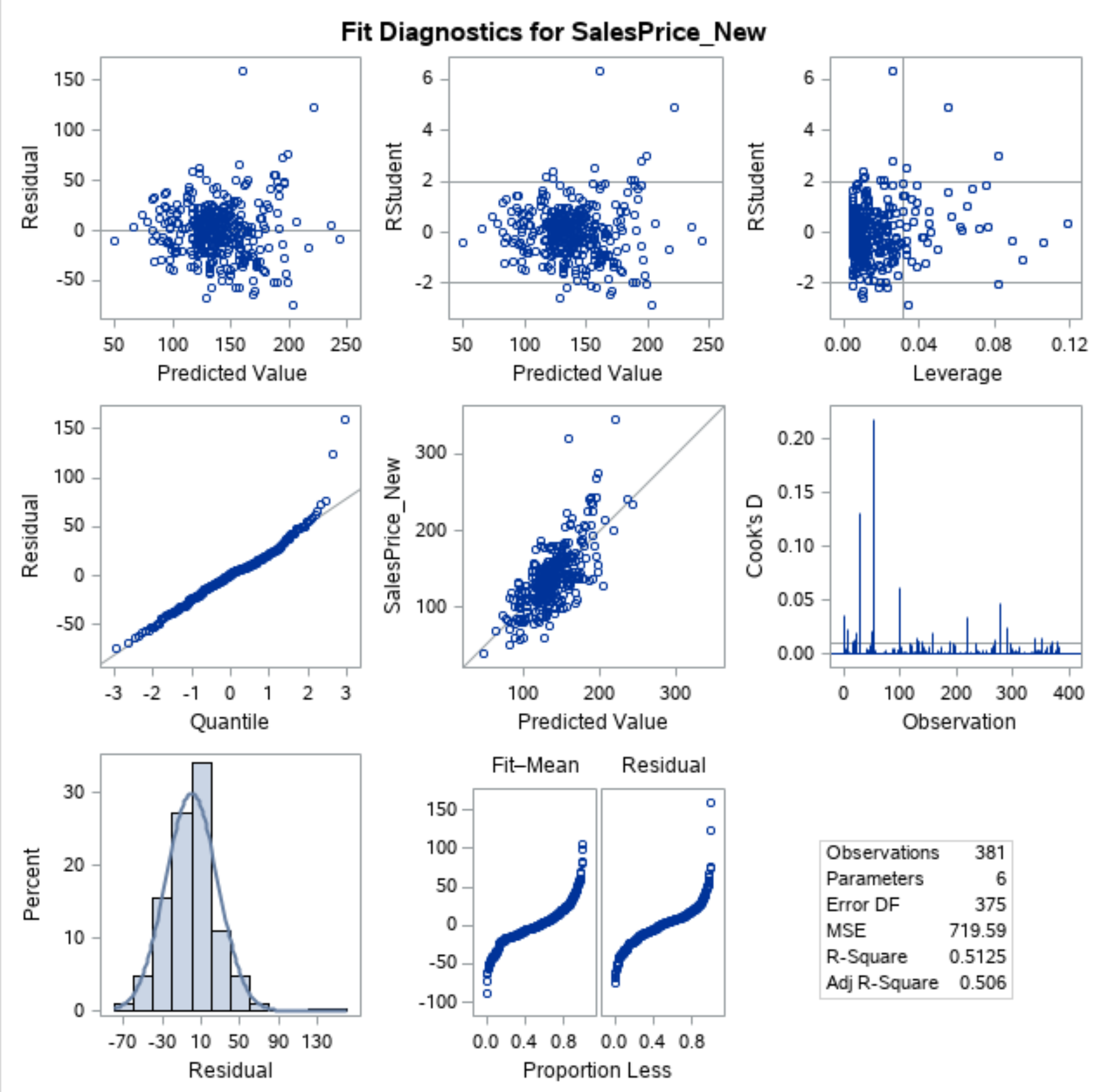
3

4

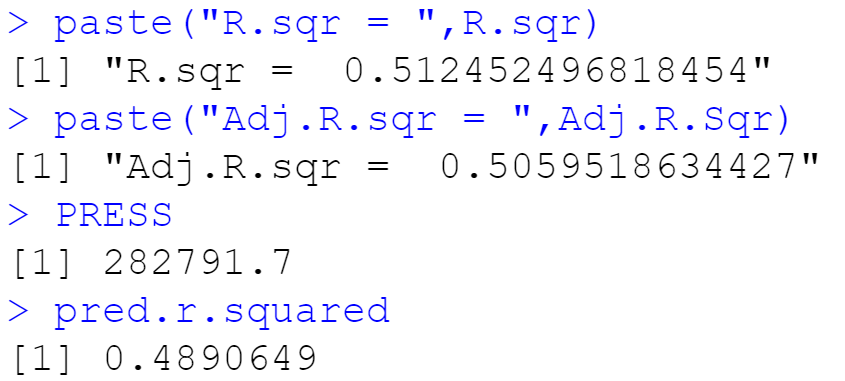
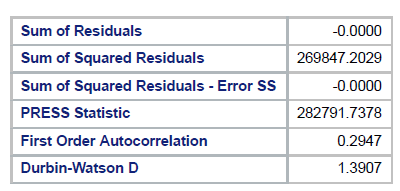
Appendix - Figure 1: Initial Plot - Sales prices v/s. Ground Living Area

Appendix - Table : Analysis1, Outlier details and anticipated reason for outliers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Outlier Details** | | | | | **Anticipated Reason** |
| **Outlier** | **Id** | **Neighborhood** | **GrLivArea  (sq. Ft)** | **SalePrice ($)** |
| 1 | 1299 | Edwards | 5642 | 160000 | The sale condition given as partial which suggest home was not completed when last assessed (associated with new homes) This might be one reason other than flattening in sale price after crossing certain value of ground living area. |
| 2 | 524 | Edwards | 4676 | 184750 | The sale condition given as partial which suggest home was not completed when last assessed (associated with new homes) This might be one reason other than flattening in sale price after crossing certain value of ground living area. |
| 3 | 725 | Edwards | 1698 | 320000 | The combination of Factors: LotArea (13286), OverallQual (9), YearBuild (2007), MasVnrArea (340), BsmtFinSF1 (1234) and  Year sold (2009) might be the reason for higher sales price |
| 4 | 643 | NAmes | 2704 | 345000 | The combination of factors: OverallQual (8),  OverallCond (7) and  Year sold (2009) might be the reason for higher sales price |



Appendix - Figure 2: Residual Plots for Interaction Model (Analysis 1)

Appendix - Figure 3: Interaction Model comparison statistics (Analysis 1)