Assignment #1

Sameer Khan

Introduction:

In this assignment, we examine a dataset that contains information on residential single family homes. The initial dataset has been narrowed down to focus on six main variables: total square feet (TotalSqftCalc), number of bathrooms (TotalBathCalc), home quality (QualityIndex), number of rooms above ground (TotRmsAbvGrd), overall home quality (OverallQual), and home condition (OverallCond). We want to determine the interaction of these predictors with a single response variable representing the selling price of a given home (SalePrice). The goal of this report is to find out which predictor variables have the most impact on sale price, the shape of the relationships between the predictors and response variable, and to confirm that sale price is the best response for analysis.

These questions are answered using three different techniques. First, correlations between each variable are established to see whether a predictor has a positive or negative effect on sale price, and how strong the effect is. Next, scatterplots are used to examine specific relationships between a single predictor variable and the response variable. Scatterplots are also shown using loess smoother to normalize the data and account for any possible skews. Finally, boxplots are created for three predictor variables to determine whether they should be treated as discrete or continuous.

Basic Correlation Analysis:

The figures shown below display correlation matrices that show the type and strength of relationship between six predictor variables and the response variable. Figure 1 does this for the chosen response variable, SalePrice. We see that five of the variables have positive relationships with SalePrice and one (OverallCond) has a negative relationship. None of the correlation coefficients are near zero which implies some relative statistical significance. The three strongest correlation coefficients with respect to SalePrice are OverallQual (0.82), TotalSqftCalc (0.79), and TotalBathCalc (0.67).

Figure 2 produces a similar correlation matrix; however, the response variable has been changed to the log of SalePrice. This is done to normalize the data given that each variable has a different unit of measurement. For example, a home's total square feet will typically be in the thousands; our data set ranges from 825 to 5771 square feet. To contrast, the total number of bathrooms or rooms will most likely be under 10. Performing a log transformation will smooth this discrepancy out and help us see if any of the original correlations were skewed towards a certain predictor. The log correlation matrix still shows a positive relationship between five variables and one negative relationship. The values of the coefficients slightly differ but the three strongest correlations remain the same.

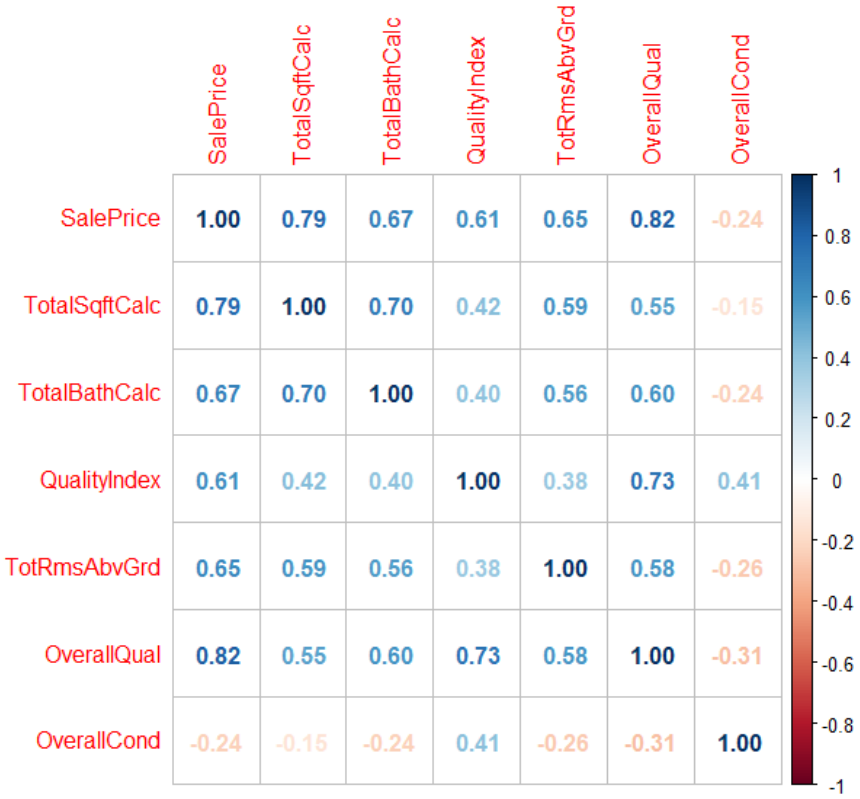


Figure 1: Correlation Matrix with SalePrice as the response

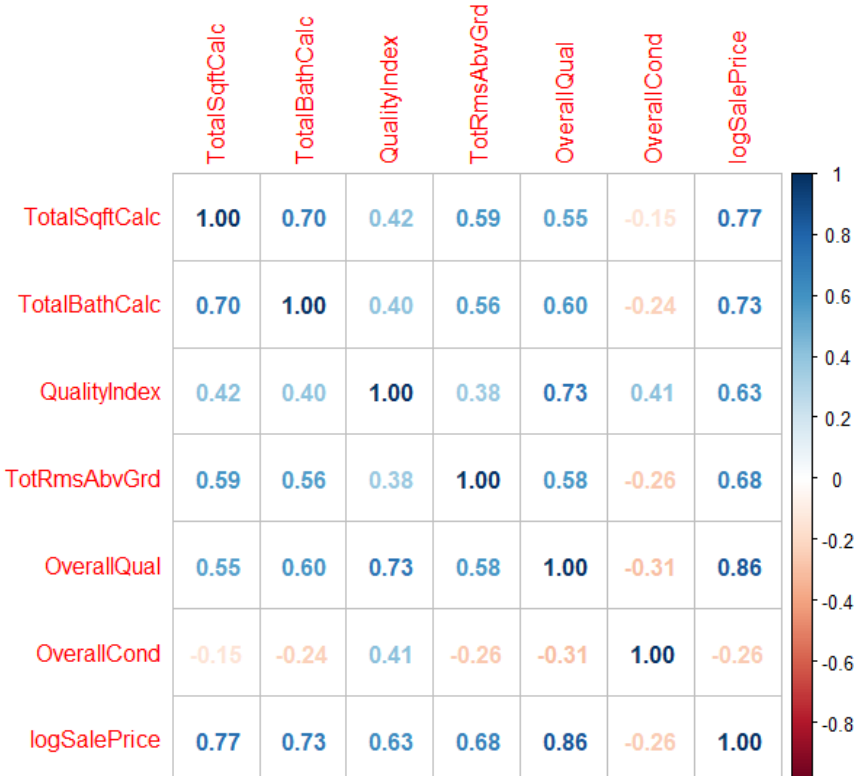
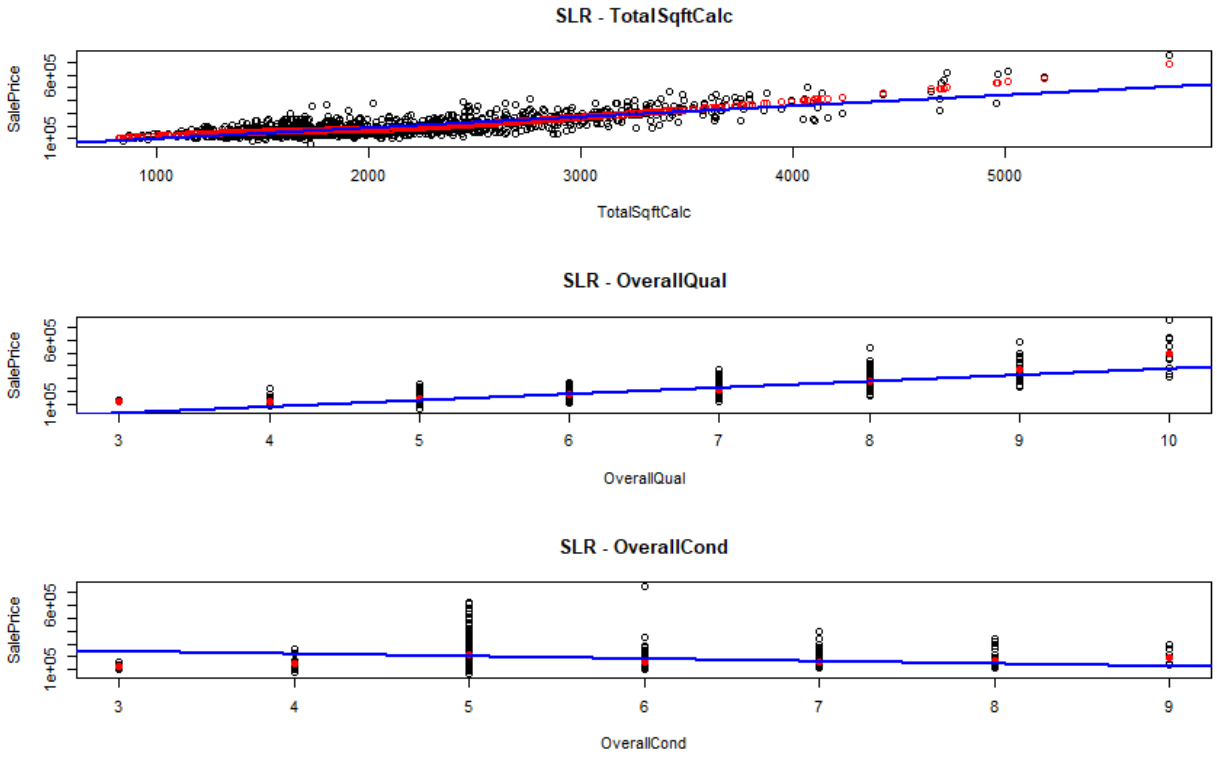


Figure 2: Correlation Matrix with log(SalePrice) as the response

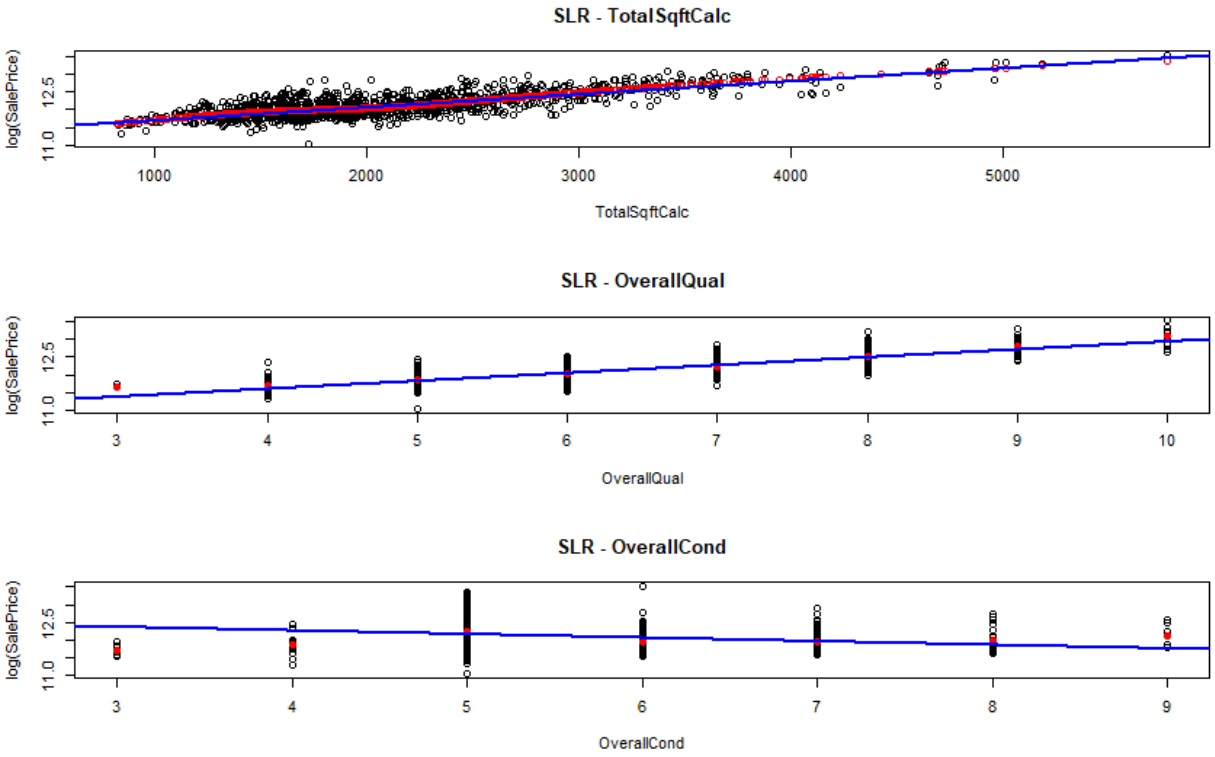
Which Response Variable?

Scatterplots have been created for three variables to show each one's relationship with SalePrice and log(SalePrice). Total square feet and overall quality were selected because they have the strongest positive correlation with SalePrice. Overall home condition is the only variable that has a negative correlation with SalePrice which makes it worth examining. Scatterplot 1 shows a positive relationship between total square feet and sale price. The loess smoother helps to condense the data and show a non-linear collection of points, and the fitted SLR shows a similar linear relationship. It is important to note that scatterplots 2 and 3 have variables that are categorical; the quality and condition respectively are recorded on a 1-10 scale. The smoother is displayed in these charts as a singular point for each value on the scale. The trend lines confirm the correlation matrix relationships.

Similar scatterplots (4-6) are created for the same variables with respect to log(SalePrice). The output is almost identical in shape and form of scatterplots 1-3. This is a good indication that the variables we selected are important in determining the selling price of a home. The log transformation cleans up each plot by reducing the spread of the data. The scatterplots for OverallCond show that as house condition improves, there is a general decrease in price, but this effect could be due to the number of houses sold at each condition point. Most of the data falls between a condition of 5 and 7, with fewer homes classified as either being in poor or excellent condition.



Scatterplots 1-3: Predictor variables with SalePrice as the response



Scatterplots 4-6: Predictor variables with log(SalePrice) as the response

Discrete or Continuous?

Figure 3 contains scatter and box plots for the variables TotRmsAbvGrd, OverallQual, and OverallCond. We can see possible outlier points using the boxplots, as well as the interquartile range for each predictor. The trend lines for Rooms above ground and quality seem to match the median values for their respective boxplots, while the trendline and median values for home condition differ in shape. I think it would be appropriate to categorize TotRmsAbvGrd as a discrete variable and OverallQual/OverallCond as continuous. The reasoning for this classification is due to the inherent meaning of each variable. Total rooms above ground will always be an integer; it's impossible to have a house with only part of a room. Each room is treated as a discrete entity, so it follows that the variable should be classified the same way. Quality and Condition are more subjective and should be made continuous. They represent a scale that has been quantified and should not be limited to singular values.

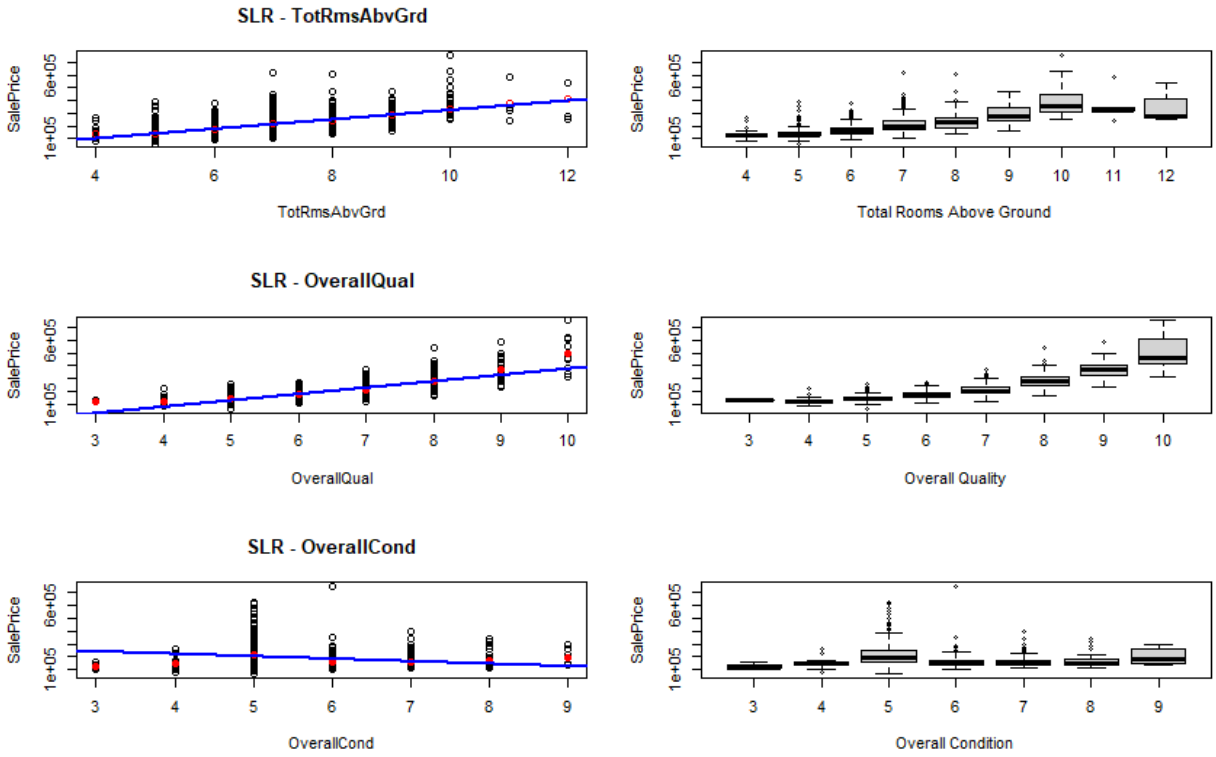


Figure 3: Scatter and Box plots for 3 predictor variables