**Module 1 Assignment Regression Diagnostics with R**

**(Ames Housing Dataset)**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

[**ALY6015.71591.202515**](https://northeastern.instructure.com/courses/196378) **: Intermediate Analytics**

**Professor: Roy Wada**

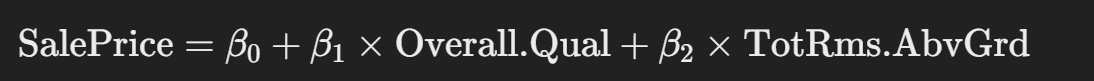
**November 3rd, 2024**

**Introduction**

In this analysis, I explored the Ames Housing dataset with the aim of identifying significant factors influencing house sale prices. I used a systematic approach involving data preparation, exploratory data analysis, regression model fitting, diagnostic checks, and model refinement. The report details each step and provides findings supported by visual representations and model interpretations.

**Analysis**

1. **Data Preparation and Handling Missing Values**  
   After loading the dataset, I addressed missing values by imputing the mean for numeric variables and the mode for categorical variables. This step ensured that all variables were available for modeling (Appendix **A.1**).
2. **Exploratory Data Analysis (EDA)**  
   I created descriptive statistics for SalePrice, along with a histogram and a box plot to understand its distribution. The histogram (Appendix **A.2**) indicated a right-skewed distribution, suggesting that most houses fall within a lower price range, with some high-price outliers. The box plot (Appendix **A.3**) visually confirmed these outliers.
3. **Correlation Analysis**  
   A correlation matrix (Appendix **A.4**) was generated to examine relationships between numeric variables. This analysis showed that Overall.Qual (overall quality) had the highest positive correlation with SalePrice, PID had the lowest correlation, and TotRms.AbvGrd (total rooms above grade) showed a moderate correlation. I also created scatter plots to illustrate these relationships:
   * **Overall.Qual vs. SalePrice** (Appendix **A.5**)
   * **PID vs. SalePrice** (Appendix **A.6**)
   * **TotRms.AbvGrd vs. SalePrice** (Appendix **A.7**)
4. **Initial Regression Model**  
   I fitted an initial regression model using Overall.Qual and TotRms.AbvGrd as predictors. The model equation is as follows:



The model summary (Appendix **A.8**) revealed that both predictors had significant positive impacts on SalePrice, with an R-squared value of approximately 68%, suggesting that these predictors explain a substantial portion of the variability in SalePrice.

1. **Diagnostic Checks**  
   To validate model assumptions, I generated diagnostic plots (Appendix **A.9**). The residual plots suggested minor variance issues, but overall, the assumptions held. Variance Inflation Factor (VIF) analysis (Appendix **A.10**) confirmed no multicollinearity. Cook’s Distance plot (Appendix **A.11**) helped identify influential points, which were then removed for further refinement.
2. **Refined Model**  
   After removing influential points, I refitted the model, resulting in an improved R-squared of approximately 73% (Appendix **A.12**). This refined model confirmed similar effects of Overall.Qual and TotRms.AbvGrd on SalePrice, indicating the reliability of these predictors.
3. **Best Model Selection Using All-Subsets Regression**  
   Using all-subsets regression, I evaluated models with different predictor combinations. The best model based on Bayesian Information Criterion (BIC) included Overall.Qual, TotRms.AbvGrd, Lot.Area, and Year.Built and can be represented as:

When comparing AIC values, the refined model showed a better fit, balancing simplicity and performance (Appendix A.13).

**Conclusion/Interpretations**

The analysis identified Overall.Qual and TotRms.AbvGrd as significant predictors of SalePrice. The refined model, with high R-squared and low AIC values after removing influential observations, provided an optimal fit. The final model indicates that higher overall quality and more rooms above grade contribute positively to sale prices. These findings provide valuable insights for real estate pricing strategies and offer guidance to buyers and sellers on factors that affect home valuation.

**References**

Ames Housing Dataset Documentation. Provided by Professor. (n.d.). (Accessed: November 1, 2024).

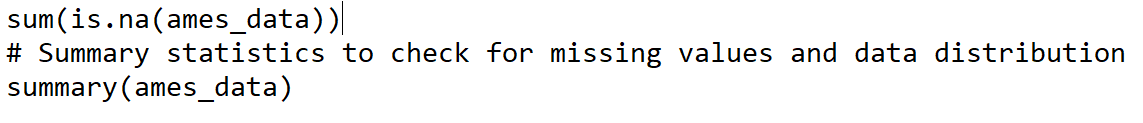
R Documentation. (n.d.). lm function. Available at: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm> (Accessed: November 1, 2024).

ggplot2 Documentation. (n.d.). ggplot2 package in R. Available at: <https://ggplot2.tidyverse.org/> (Accessed: November 2, 2024).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer. Available at: <https://www.statlearning.com/> (Accessed: November 2, 2024).

**Appendix**

* **A.1**: Data Summary and Missing Values  
  *Description*: Provides summary statistics and missing value counts for the dataset.



A screenshot of a computer

Description automatically generated

* **A.2**: Histogram of SalePrice  
  *Description*: Histogram showing the distribution of SalePrice.

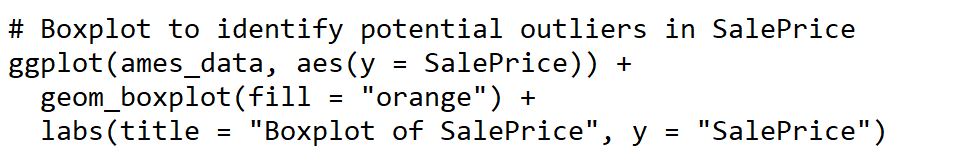
A computer code with black text

Description automatically generated with medium confidence

A graph of a distribution of sales

Description automatically generated

* **A.3**: Box Plot of SalePrice  
  *Description*: Box plot highlighting outliers in SalePrice.



A graph with a line and a line

Description automatically generated with medium confidence

* **A.4**: Correlation Matrix of Numeric Variables  
  *Description*: Correlation matrix identifying relationships between variables.

A screenshot of a computer code

Description automatically generated

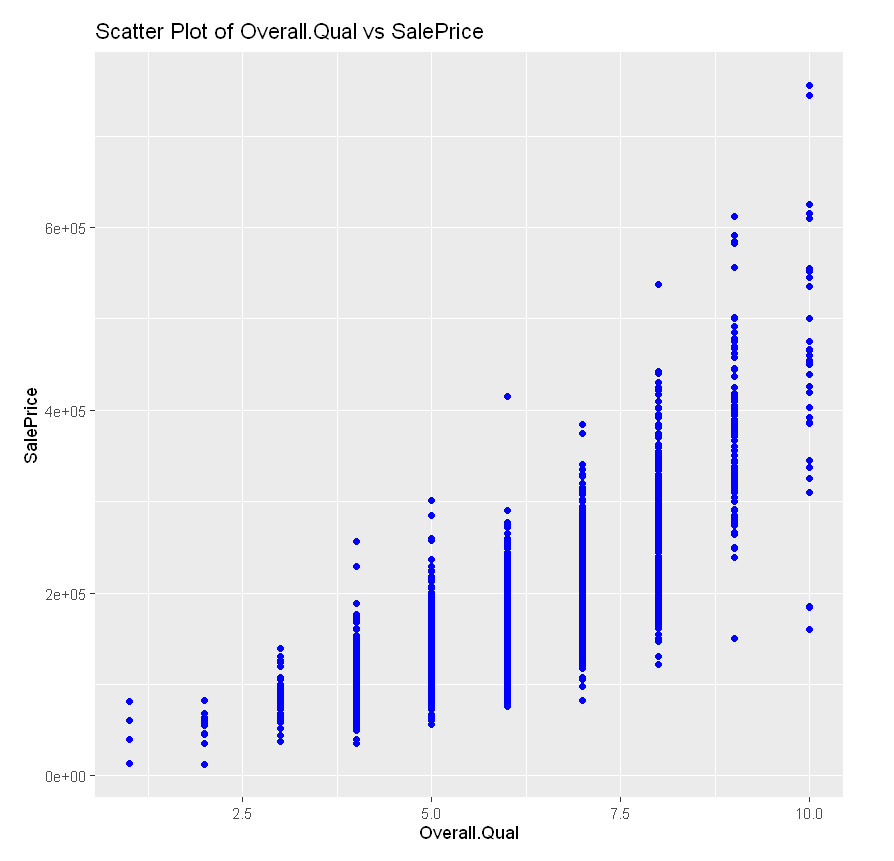
A diagram with blue and red dots

Description automatically generated

* **A.5**: Scatter Plot of Overall.Qual vs. SalePrice  
  *Description*: Scatter plot showing the positive relationship between Overall.Qual and SalePrice.

A close up of text

Description automatically generated



* **A.6**: Scatter Plot of PID vs. SalePrice  
  *Description*: Scatter plot showing no clear relationship between PID and SalePrice.

A close up of a text

Description automatically generated

A graph with red dots

Description automatically generated

* **A.7**: Scatter Plot of TotRms.AbvGrd vs. SalePrice  
  *Description*: Scatter plot showing a moderate positive relationship between TotRms.AbvGrd and SalePrice.

A close-up of a white background

Description automatically generated

A graph of green dots

Description automatically generated

* **A.8**: Initial Model Summary  
  *Description*: Summary of the initial regression model with Overall.Qual and TotRms.AbvGrd as predictors.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

* **A.9**: Diagnostic Residual Plots  
  *Description*: Residual plots for evaluating model assumptions.

A white background with black text

Description automatically generated

A group of graphs and diagrams

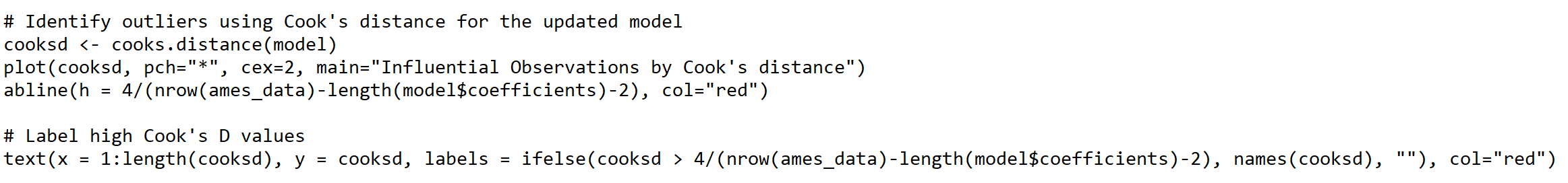
Description automatically generated

* **A.10**: VIF for Initial Model  
  *Description*: VIF values for multicollinearity check in the initial model.

A close-up of a sign

Description automatically generated

* **A.11**: Cook’s Distance Plot  
  *Description*: Cook’s Distance plot for identifying influential observations.



A graph with red dots

Description automatically generated

* **A.12**: Refined Model Summary  
  *Description*: Summary of the refined regression model after removing influential observations.

A close-up of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

* **A.13**: Best Model Summary from All-Subsets Regression  
  *Description*: Summary of the best subset model identified by BIC.

A white background with black text

Description automatically generated

