Regression Diagnostic and Transformation Report

1. OLS Simple Regression Model

1.1 Variable Selection

Among the continuous variables in the *Ames Housing* dataset, *GrLivArea* was selected as the predictor to build a simple OLS regression model with the response variable *SalePrice*, labeled Model 1. This variable was selected after conducting a correlation matrix between all continuous variables. The correlation coefficient between *GrLivArea* and *SalePrice* was 0.7067, the second highest only below *OverallQual*. This indicates a strong linear relationship highlighting its useful for the model. Additionally, *GrLivArea* represents a quantitative and objective measure of housing sale prices as it describes the total living space in a home, a key attribute that buyers look for in a home.

1.2 Scatterplot of SalePrice

A scatterplot with the regression line of *SalePrice* and *GrLivArea* was performed to assess the relationship between the variables. The data is roughly around the regression line, indicated good fit of model. Moreover, the regression line indicates a positive linear relationship with *SalePrice* and *GrLivArea*, with the larger sized homes having higher sale prices. This is intuitive with the variables selected, where more living space equates to a higher prices home.

1.3 Fitted Model Equation

A simple OLS regression model was fitted with *SalePrice* (Y) and *GrLivArea* (X). The fitted equation for the model can be seen below:

$$\boldsymbol{\hat{Y} = 27516.45 + 97.16 * (GrLivArea)}$$

According to table 1 below, each coefficient, including the model intercept, was statistically significant $(p < 0.001)$. The model intercept was statistically significant ($t = 9.486$, $p < 0.001$), indicates that when *GrLivArea* is zero, the *SalePrice* of a home is $27,516.45. Though statistically meaningful, practically, we cannot use this information as a home with zero square feet of living space is unrealistic. The *GrLivArea* slope coefficient, 97.16 indicates for every 1-unit increase in *GrLivArea*, the *SalePrice* increased on average by $97.16.

Table 1. Coefficient Table for Model 1

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	80139.794	3040.750	-26.36	<2e-16 ***
GrLivArea	55.882	1.605	34.81	<2e-16 ***

The null hypothesis (H_0) of the coefficients states that slope is zero, suggesting no linear relationship between the variables; the alternative hypothesis (H_1) of the coefficients states that the slope is nonzero, suggesting there is a linear relationship between variables. The hypotheses can be written as:

$$\mathbf{H_0: \beta_1 = 0}$$
$$\mathbf{H_1: \beta_1 \neq 0}$$

The slope coefficient for *GrLivArea* was statistically significant, indicating that we can *reject* the null hypothesis and conclude that there is a linear relationship between *GrLivArea* and *SalePrice*.

1.4 R^2 Interpretation

The coefficient of determination, R^2 value, for Model 1 was 0.5003, indicating that roughly 50.03% of the variability in *SalePrice* is explained by *GrLivArea*. This signifies roughly 50% of the variation in *SalePrice* that remains unexplained, suggesting the need for adding more meaningful variables to increase the R^2 value and explain more of the variation in *SalePrice*.

1.5 Model Analysis

An omnibus F-test was performed to determine if the overall model was significant and if the model explains a significant amount of variance in response variable, *SalePrice*. The null hypothesis (H_0) for the omnibus F-test states that the slope of the coefficient is zero, indicating no relationship between the variables, while the alternative hypothesis (H_1) states that the slope of the coefficient is nonzero, meaning there is a relationship between the variables. The hypotheses can be written as:

$$\mathbf{H_0: \beta_1 = 0}$$
$$\mathbf{H_1: \beta_1 \neq 0}$$

According to table 2, the omnibus F-test was statistically significant ($F_{(1,2653)} = 2656, p < 0.001$), suggesting that *GrLivArea* has a significant linear relationship with *SalePrice*. Additionally, the *GrLivArea* coefficient was statistically significant ($F = 2655.9, p < 0.001$), indicating that *GrLivArea* is a strong predictor and has a linear relationship with *SalePrice*.

Table 2. ANOVA Table for Model 1

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr (>F)**
GrLivArea	1	5.5910e+12	5.5910e+12	2655.9	<2e-16
Residuals	2653	5.5849e+12	2.1051e+09		
Model	1, 2653			2656	< 2.2e-16

1.6 Model Assumption Analysis

The hypothesis tests for the coefficients and the overall model are dependent on normality, independence, and homoscedasticity assumptions being met. To assess normality, a histogram of the standardized residuals was conducted. The histogram indicates that the standardized residuals are normally distributed around zero. However, there are observations spread to from -2.5 to over

+5 standard deviations, pointing to significant outliers and moderate right skewness ultimately violating the normality assumption.

There is no plot to measure independence of the variables, however, the variables are likely independent because the dataset represents different properties sold at various times, there are no repeated measures on the same homes, and there is no obvious clustering between the variables. In section 1.1, the selected variables had a strong positive correlation coefficient of 0.767, which signifies that *GrLivArea* is a good predictor of *SalePrice*.

Homoscedasticity checks that the residuals maintain constant variance across all independent variables. To assess homoscedasticity, a scatterplot of the standardized residuals and predicted values of model was performed, which provides strong evidence of heteroscedasticity. This is seen as a cone shaped spread of the residuals increases systemically as the predicted values increase. As the predicted values increase towards $200,00-$250,000, the vertical spread of residuals becomes noticeably wider.

1.7 Model Diagnosis Analysis

Important diagnostic measures such as leverage, influence, and outlier tests help evaluate the appropriateness of the regression model using *GrLivArea* as a predictor to *SalePrice*. These diagnostic measures provide statistical support for the underlying assumptions of an OLS regression model and can uncover problems in the model. Figure A1 provides visualizations of common diagnostic measures like leverage, outlier tests, and Cook's Distance (influence) for Model 1. Figure A2 is a bubble plot of potential influential observations in Model 1.

The hat-value plot in Figure A1, and the bubble chart reveal several observations with high leverage, meaning they are extreme *GrLivArea* values far from the mean. Notably, observations 1417, 2022, and 2486 show hat-values between 0.006 to 0.010, exceeding the threshold of 0.0015 for Model 1. These observations represent homes with living areas at the extremes of the data distribution. While having high leverage is not inherently an issue, it becomes concerning when it is combined with large residuals, creating "bad leverage points."

The Cook's Distance measures in Figure A1, and the bubble chart reveals critical influence problems in Model 1. Observation 2022 was the most visible and problematic, with Cook's D of 0.0398, well above the threshold of 0.0015 for Model 1. Other notable observations include 1417, which also had a Cook's D greater than 0.03. This concentration of influence in these observations means that Model 1's coefficients, standard errors, and predications are unreliable.

The Studentized residuals plot in Figure A1 identifies multiple potential extreme outliers in the model. Observations 1595 and 2184 show standardized residuals exceeding +4.0, well beyond the ±3 threshold for extreme outliers. These outliers represent where the linear model severely underpredicts the actual *GrLivArea* value. The histogram of the standardized residuals, seen in Figure A1, confirms this pattern seen in the long right tail extending to +5, suggesting that the model consistently underpredicting certain high value properties more severely than it overpredicts low-value ones.

To correct these issues, the data points will need to be examined to determine if they are valid entries. If they are found to be accurate, then these points can either be removed or a logarithmic transformation of *SalePrice* can reduce influence, heteroscedasticity, and improve normality. Additionally, more predictive variables can be added to the model to better explain the variance seen in *SalePrice* that *GrLivArea* is not capturing. After implementing changes, the diagnostic plots must be reassessed to verify the model meets regression assumptions.

2. OLS MLR Model

2.1 Fitted Model Equation

A multilinear regression (MLR) model was fitted with *GrLivArea* (X1) and *OverallQual* (X2) as the predictor variables to *SalePrice* (Y) labeled Model 2. The fitted equation of the fitted MLR model was:

$$\boldsymbol{\hat{Y} = -80139.79 + 55.88 * (GrLivArea) + 28339.49 * (OverallQual)}$$

According to table 3 below, each coefficient including the intercept, was statistically significant *($p < 0.001$)*. The *GrLivArea* slope coefficient indicates for every 1-unit increase in *GrLivArea*, the *SalePrice* increased on average by $55.88 when controlling for *OverallQual*. Additionally, the *OverallQual* slope coefficient suggests for every 1-unit increase in *OverallQual*, the *SalePrice* increased on average by $28,339.49 when accounting for GrLivArea. The difference between this fitted equation to the fitted equation in section 1.3, each coefficient represents a partial effect.

Table 3. Coefficient Table for Model 3

Coefficients	**Estimate**	**Std. Error**	**t-value**	**Pr (>|t|)**
Intercept	80139.794	3040.750	-26.36	<2e-16
GrLivArea	55.882	1.605	34.81	<2e-16
OverallQual	28339.48	578.400	49.00	<2e-16

The null hypothesis (H_0) of the coefficients states that all slope values are zero, suggesting no linear relationship between the explanatory and response variables; the alternative hypothesis (H_1) of the coefficients states that all slope values are nonzero, suggesting there is a linear relationship between the explanatory and response variables. The hypotheses can be written as:

$$\mathbf{H_0: \beta_1 = \beta_2 = 0}$$

$$\mathbf{H_1: \beta_1 \neq \beta_2 \neq 0}$$

The slopes of *GrLivArea* and *OverallQual* were statistically significant, suggesting we can *reject* the null hypothesis and conclude that both variables have a linear relationship with *SalePrice*.

2.2 R^2 Interpretation

The coefficient of determination, R^2 value, for Model 2 was 0.7377, indicating that roughly 73.77% of the variability in *SalePrice* is explained by *GrLivArea* and *OverallQual*. This signifies roughly 23% of the variation in *SalePrice* that remains unexplained, suggesting Model 2 provides a significant amount of explanation of the variation in *SalePrice*. Moreover, the R^2

value increased by 0.24 when compared to Model 1, indicating Model 2 is a better model to characterize *SalePrice*.

2.3 Model Analysis

An omnibus F-test was performed to determine if the overall model was statistically significance and if the model explains a significant amount of variance in SalePrice. The null hypothesis (H_0) for the omnibus F-test states that the slopes of the coefficients are zero, indicating no relationship between the variables, while the alternative hypothesis (H_1) states that the slopes of the coefficients are nonzero, meaning there is a relationship between the variables. The hypotheses can be written as:

$$\mathbf{H_0: \beta_1 = \beta_2 = 0}$$

$$\mathbf{H_1: \beta_1 \neq \beta_2 \neq 0}$$

According to table 4 below, the omnibus F-test was statistically significant ($F_{(2,2652)} = 3729, p < 0.001$), suggesting that *GrLivArea* and *OverallQual* have a significant linear relationship with *SalePrice*. Additionally, both the *GrLivArea* coefficient ($F = 5058.2, p < 0.001$) and *OverallQual* coefficient ($F = 2400.6, p < 0.001$) were statistically significant, indicating that both variables are a strong predictor and have a linear relationship with *SalePrice*. *GrLivArea* had a substantially larger F-statistic, suggesting that *GrLivArea* is a better predictor of the *SalePrice* than *OverallQual* and explains more of the variance in *SalePrice*.

Table 4. ANOVA Table for Model 2

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr (>F)**
GrLivArea	1	5.5910e+12	5.5910e+12	5058.2	<2e-16
OverallQual	1	2.6535e+12	2.6535e+12	2400.6	<2e-16
Residuals	2652	2.9314e+12	1.1053e+09		
Model	2, 2652			3729	<2e-16

2.4 Model Assumption Analysis

Inference regarding model coefficients and overall significance depends on meeting the assumptions of homoscedasticity, independent observations, and normality. To assess normality, a histogram of standardized residuals was conducted. The distribution of standardized residuals is symmetric distribution centered around zero with most of the observations falling between -3 and +3 standard deviations. There is a slight positive skew with a long right tail and some extreme positive outliers beyond +3, and notably, there are a few extreme negative outliers extending to -6. Despite these observations, the normality assumption is reasonably met as the distribution is approximately normal in the center.

There is no plot to measure independence of the variables, however, the variables are likely independent because the dataset represents different properties sold at various times, there are no repeated measures on the same homes, and there is no obvious clustering between the variables, signaling the independent assumption has been reasonably met.

To assess homoscedasticity, a scatterplot of the standardized residuals and predicted values of model was performed, which provides evidence of heteroscedasticity. This is seen as a cone shaped spread of the residuals increases as the predicted values increase. As the predicted values increase towards higher ends ($200,00-$250,000), the variance progressively expands.

2.5 Model Diagnosis Analysis

Important diagnostic measures help evaluate the appropriateness of the regression model using *GrLivArea* and *OverallQual* as predictors to *SalePrice*. These diagnostic measures provide statistical support for the underlying assumptions of an OLS regression model and can uncover problems in the model. Figure A3 provides visualizations of common diagnostic measures like leverage, outlier tests, and Cook's Distance (influence) for Model 2. Figure A4 is a bubble plot of potential influential observations in Model 2.

The hat value plot in Figure A3 and the bubble chart highlight a few observations with high leverage. Observation 1417 has particularly high leverage, hat-value of 0.016, well over the threshold of 0.0015 for Model 2, but low influence because its residual is small. Additionally, observation 2486 has the second largest hat-value of 0.012 and had a large positive residual value of 4 indicating high leverage.

The Cooks Distance chart in Figure A7 and the bubble chart reveal several points with high influence. Observation 2486 had the highest Cook's D value of 0.052, above the threshold of 0.0015 for Model 2. With the high leverage of observation 2486 and large negative residual, the high influence indicates this observation will need to be fixed and is especially problematic. Other notable observations include 1107 (Cook's D = 0.047, hat-value = 0.0043) and 1553 (Cook's D = 0.015, hat-value = 0.0016) indicating high influence but moderate leverage.

Outlier tests in the Studentized residuals plots in Figure A7 indicate several concerning observations beyond ±3. Record 1107 has a very low negative residual value of -5.47, 1553 has a very high positive residual value of 5.16, and 1242 has a low negative residual value of -4.73, indicating that these observations do not fit the model well.

2.6 Model 2 Conclusions

In conclusion, *GrLivArea* and *OverallQual* should be retained as predictor variables to *SalePrice*. Both variables are fundamentally important for predicting home prices because *GrLivArea* captures the size of the home and *OverallQual* captures condition and quality of the home, which represent different attributions of housing value. Moreover, both variables measure distinct aspects as living area is objective and quantitative, while overall quality is subjective and qualitative. Together, these variables provide a more complete picture of housing sales than either variable could alone. The influential points and outliers observed in the diagnostic plots do not suggest problems with the predictor variables themselves, rather that the model is underspecified and we need to add more predictors than removing the existing ones. Like the simple regression model, a transformation to the response variable (*SalePrice*) may assist with addressing the heteroscedasticity observed in the diagnostic criteria.

3. Model 3 Analysis

3.1 Fitted Model Equation

A multilinear regression (MLR) was fitted with *GrLivArea* (X1), *OverallQual* (X2), and *YearBuilt* (X3) as the predictor variables to *SalePrice* (Y), called Model 3. The fitted equation of the fitted MLR model was:

$$\boldsymbol{\widehat{Y} = -1066000 + 60.11(GrLivArea) + 20990(OverallQual) + 5202(YearBuilt)}$$

According to the coefficient table seen below in Table 5, each coefficient was statistically significant, indicating that, individually, when each variable is tested against *SalePrice*, it has a positive linear relationship. Specifically, each estimate is a partial effect, meaning the amount of increase seen in *SalePrice* is associated when all other variables in the models are held constant. For instance, with 1 square foot in GrLivArea, the sale price of a home differed by $60.11 but have the same quality rating and were built in the same year. This allows for a more refined comparison of the variables by removing cofounding variables.

Table 5. Coefficient Table for Model 5

Coefficients	**Estimate**	**Std. Error**	**t-value**	**Pr (>|t|)**
Intercept	-1066000	46660	-22.86	<2e-16
GrLivArea	60.11	1.498	40.12	<2e-16
OverallQual	20990	637.7	32.91	<2e-16
YearBuilt	520.2	24.56	21.18	<2e-16

3.2 R^2 Interpretation

The coefficient of determination (R^2 value) for Model 3 was 0.7757, indicating that 77.57% of the variation in *SalePrice* is explained by *GrLivArea*, *OverallQual*, and *YearBuilt*. This suggests that there was an increase in R^2 value from Model 2 of 0.0380, meaning that *YearBuilt* added an additional 3.80% variance in *SalePrice* that was not explained by *GrLivArea* and *OverallQual* cannot explain alone. The resulting difference in R^2 value signifies that *YearBuilt* improved the models explanatory ability since the coefficient for *YearBuilt* was highly statistically significant with a large test statistic, providing evidence that the relationship between *YearBuilt* and *SalePrice* is real and not due to chance. Additionally, the residual standard error (RSE) reduced from $33,250 in Model 2 to $30,750 in Model 3. The reduction in RSE signifies that the predictions in Model 3 are $2,500 closer to the true sale prices than Model 2.

3.3 Coefficient & Overall Model Analysis

An omnibus F-test was performed to determine if the overall model was statistically significance and if the model explains a significant amount of variance in SalePrice. The null hypothesis (H_0) for the omnibus F-test states that the slopes of the coefficients are zero, indicating no relationship between the variables, while the alternative hypothesis (H_1) states that the slopes of the coefficients are nonzero, meaning there is a relationship between the variables. The hypotheses can be written as:

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_1: $\beta_1 \neq \beta_2 \neq \beta_3 \neq 0$

According to table 4 below, the omnibus F-test was statistically significant ($F_{(3,2651)} = 3055, p < 0.001$), suggesting that *GrLivArea*, *OverallQual*, *YearBuilt* have a significant linear relationship with *SalePrice*. Additionally, *GrLivArea* coefficient, *OverallQual*, and YearBuilt were statistically significant, indicating that all variables are a strong predictor and have a linear relationship with *SalePrice*. *GrLivArea* had the largest F-statistic, suggesting that *GrLivArea* is a better predictor of the *SalePrice* than *OverallQual* and *YearBuilt* variables.

Table 6. ANOVA Table for Model 3

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr (>F)**
GrLivArea	1	5.5910e+12	5.5910e+12	5911.72	< 2.2e-16
OverallQual	1	2.6535e+12	2.6535e+12	2805.75	< 2.2e-16
YearBuilt	1	4.2419e+11	4.2419e+11	448.52	< 2.2e-16
Residuals	2651	2.5072e+12	9.4575e+08		
Model	3, 2651			3055	< 2.2e-16

3.4 Model Assumption Analysis

Inference regarding model coefficients and overall significance depends on meeting the assumptions of homoscedasticity, independent observations, and normality. To assess normality, a histogram of standardized residuals was conducted. The distribution of standardized residuals is symmetric distribution centered around zero with most of the observations falling between ±3 standard deviations. There is a slight positive skew with a long right tail and some extreme positive outliers beyond ±3, and notably, there are a few extreme negative outliers extending to ±6. Despite these observations, the normality assumption seems to be reasonably well met.

There is no plot to measure independence of the variables, however, the variables are likely independent because the dataset represents different properties sold at various times, there are no repeated measures on the same homes, and there is no obvious clustering between the variables, signaling the independent assumption has been reasonably met.

Like other models above, homoscedasticity was found to be violated when plotting the predicted values versus the standardized residuals. A cone shaped spread of the residuals increases as the predicted values increase. As the predicted values increase towards higher ends ($200,00-$250,000), the variance progressively expands. The heteroscedasticity pattern similar to Model 2 suggests that *YearBuilt* did not resolve the variance issue.

3.5 Model Diagnosis Analysis

Important diagnostic measures help evaluate the appropriateness of the regression model using *GrLivArea* and *OverallQual* as predictors to *SalePrice*. These diagnostic measures provide statistical support for the underlying assumptions of an OLS regression model and can uncover problems in the model. Figure A5 provides visualizations of common diagnostic measures like leverage, outlier tests, and Cook's Distance (influence) for Model 3. Figure A6 is a bubble plot of potential influential observations in Model 3.

The hat value plot in Figure A5 and the bubble chart highlight a few observations with high leverage. Observations 1417 and 1244 has particularly high leverage, hat-values over the threshold of 0.0015 for Model 3, but low influence because its residual is small. Observation 2486 reported a large hat-value and had a large positive residual value of 4 indicating high leverage.

The Cooks Distance chart in Figure A5 and the bubble chart reveal several points with high influence. Observation 2486 had the highest Cook's D value of 0.052, above the threshold of 0.0015 for Model 2. With the high leverage of observation 2486 and large negative residual, the high influence indicates this observation will need to be fixed and is especially problematic. Other notable observations include 1107 (Cook's D = 0.039, hat-value = 0.0046) and 1553 (Cook's D = 0.013, hat-value = 0.0018) indicating high influence but moderate leverage.

Outlier tests in the Studentized residuals plots in Figure A5 indicate several concerning observations beyond ±3. Record 1107 has a very low negative residual value of -5.83, while 1553 has a very high positive residual value of 5.132, indicating that these observations do not fit the model well.

3.6 Model 3 Conclusions

In conclusion, *GrLivArea*, *OverallQual*, and *YearBuilt* should be retained as predictor variables to *SalePrice*. All three variables are fundamentally important for predicting home prices because *GrLivArea* captures the size of the home, *OverallQual* captures condition and quality of the home, and *YearBuilt* characterizes the age and modernity of the home, which together represent different attributions of housing value. Together, these variables provide a more complete picture of housing sales than either variable could alone. The influential points and outliers observed in the diagnostic plots do not suggest problems with the predictor variables themselves, rather that the model is underspecified and we need to add more predictors than removing the existing ones. Like the simple regression model, a transformation to the response variable (*SalePrice*) may assist with addressing the heteroscedasticity observed in the diagnostic criteria.

4. Model 4 Analysis

To fit Model 4, a logarithmic transformation was applied to *SalePrice*, aptly named *LogSalePrice*. The model was refit using the same explanatory variables as Model 3, which included *GrLivArea*, *OverallQual*, and *YearBuilt* with the transformed *LogSalePrice* as the response variable. Model 4 had a coefficient of determination value of 0.7793, indicating a very small improvement from Model 3 and explains slightly more variance in the response variable. However, the Model 4 F-statistic was 3119, which was a 64-unit increase from Model 3 of 3055, suggesting that a marginally stronger overall fit. Additionally, the scatterplot of the predicted values and the standardized residuals to assess homoscedasticity and the logarithmic transformations dramatically improved the shape of the variance, with a consistent horizontal band centered around zero and no funnel shape. This, paired with the marginal increase in explanatory power of Model 4, ultimately justifies the logarithmic transformation of the *SalePrice* variable.

5. Removing Problematic Variables

From the diagnostic plots performed of Model 4, two of the most problematic observations were removed, 169 and 1472. These observations had high influence (Cook's D between 0.08-0.14) and had extremely low residuals (> -9.0) indicating highly influential outliers. Moreover, after further investigation of those observations, the *SaleCondition* of those transactions was labels as Abnormal, indicating a non-standard transaction, and the *SalePrice* was very low, between ~$12,000-$13,000, well below the median $155,000. This suggests that the model was severely underpredicts these values as seen in the low standardized residuals. Model 4 was refitted and provided marginally better fit (R^2 = 7865, RSE = 0.1706) than the original fitted Model 4 (R^2 = 0.7793, RSE = 0.1764). Based on the marginal better fit of the new model and investigation of the data points, the justification of removing these variables is justified as they were already flagged as abnormal by the dataset itself and are not representative of the target population (i.e., typical home sales).

6. Best Model for SalePrice

To develop a comprehensive regression model predicting home sale prices in the Ames Housing dataset, I started with Model 5 and identified two continuous variables based on correlation values with LogSalePrice. TotalBsmtSF (r = 0.598) and GarageArea (r = 0.606) as promising candidates, as both showed strong correlations with price and represent distinct attributes of property value, basement space and garage capacity which both add functional living and storage space.

The selection criteria were based on increased R^2 value from Model 5, statistical significance, and VIF less than 5 to avoid multicollinearity. The R^2 value for this new model was 0.8178, providing a difference of 0.0314 from Model 5, indicating an increase in 3.14% of explanatory power over the baseline model. Both variables were highly statistically significant coefficients (p < 0.001) and low VIF (1.339 and 1.552) indicating low multicollinearity, ultimately improving predictive accuracy while avoiding overfitting.

The final model, labeled Model 6, predicts the natural log of SalePrice using five continuous predictors: above-ground living area (*GrLivArea*), overall quality rating (*OverallQual*), year built (*YearBuilt*), total basement square footage (*TotalBsmtSF*), and garage area (*GarageArea*). The model equation is:

$$\boldsymbol{\hat{Y} = 5.665 + 0.000272(GrLivArea) + 0.0102(OverallQual) + 0.00256(YearBuilt) + 0.000172(TotalBsmtSF) + 0.000218(GarageArea)}$$

All coefficients, including the intercept, are highly statistically significant *(p < 0.001*). *OverallQual* had the strongest effect, with each 1-point increase in quality rating, it raises the sale price by 10.7% ($e^{0.0102} \approx 1.107$), holding all other variables constant. *GrLivArea* shows that each additional square foot of above-ground living space increases price by 0.027%. *YearBuilt* indicates each year newer the home was built increases value by 0.256%. *TotalBsmtSF* increases price by 0.017% per square foot, while *GarageArea* increases price by 0.022% per square foot. *OverallQual* remains the dominant predictor, but the combination of size dimensions (living area, basement, and garage) collectively has substantial impact on home values.

The coefficient table shows all five predictors are highly statistically significant. *GrLivArea* has an estimated coefficient of 0.0002726 ($t = 33.35$, $p < 0.001$), *OverallQual* is estimated at 0.1020 ($t = 29.55$, $p < 0.001$), *YearBuilt* at 0.002557 ($t = 19.12$, $p < 0.001$), *TotalBsmtSF* at 0.0001719 ($t = 18.59$, $p < 0.001$), and *GarageArea* at 0.0002175 ($t = 11.40$, $p < 0.001$). The intercept is 5.665 (t = 22.30, p < 0.001). All t-values are very large, providing evidence that each value has a real relationship with *LogSalePrice* even after controlling for the other predictors.

The ANOVA table reveals the sequential contributions of each variable. When entered first, *GrLivArea* explains 175.915 sum of squares ($F = 6{,}850.15$, $p < 0.001$), making it the single largest contributor. Adding *OverallQual* next contributes 97.102 sum of squares ($F = 3{,}781.17$, $p < 0.001$), accounting for a substantial amount of variance. YearBuilt ($F= 690.31$, $p < 0.001$), TotalBsmtSF ($F= 433.17$, $p < 0.001$) and GarageArea ($F = 129.94$, $p < 0.001$), with moderately large F-statistics, suggest they only add marginal increases of variance in sale price compared to *GrLivArea* and *OverallQual*. The extraordinarily large F-statistics for all variables confirm that each contributes significantly beyond those already in the model, validating the sequential selection process.

Model 6 demonstrates exceptional goodness-of-fit, explaining 81.78% of variance in *LogSalePrice* ($R^2 = 0.8178$). This represents a substantial 3.13 improvement form Model 5 ($R^2 = 0.7865$), capturing around 14.6% of previously unexplained variance. Moreover, the adjusted R^2 value ($R^2 = 0.8175$), provides evidence that the added model complexity of *TotalBsmtSF* and *GarageArea* is justified by their explanatory power. The residual standard error decreased from 0.1706 in Model 5 to 0.1603 in Model 6, translating to a significant improvement in prediction accuracy. The overall model F-statistic was statistically significant ($F_{(5,\ 2647)} = 2337$, $p < 0.001$), signifying overwhelming proof of model significance. Additionally, multicollinearity diagnostics reveal no concerns, with all variance inflation factors well below 2.5 (*GrLivArea*: 1.540, *OverallQual*: 2.113, *YearBuilt*: 1.613, *TotalBsmtSF*: 1.399, *GarageArea*: 1.553), confirming that each predictor provides unique, non-redundant information about sale prices. The correlation matrix supports this finding, showing no predictor correlations exceed 0.56. Compared to all models developed, Model 6 represents the strongest fit achieved.

Diagnostic plots reveal that Model 6 meets nearly all regression assumptions very well. The histogram of standardized residuals shows an excellent bell-shaped, symmetric distribution centered at zero with a high, concentrated peak, indicating that the normality assumption is very well met. Most residuals fall within ±2 standard deviations, with only a few extreme outliers around -11 being the most notable. The residuals versus predicted values plot demonstrates that the homoscedasticity assumption is excellently met, showing a perfect horizontal band pattern with constant variance across all predicted values (log scale $11.00 to $13.00). This suggests that the log transformation successfully stabilized variance, ensuring that standard errors are valid, hypothesis tests have the correct Type I error rates, and predictions are equally reliable across all price ranges.

7. Summary

Variable transformation and outlier removal profoundly improved this modeling process. The logarithmic transformation of the response variable represented a critical decision, which completely removed the severe heteroscedasticity that plagued Model 3. Without this transformation, the hypothesis tests were unreliable, confidence intervals lacked proper coverage, and predictions were far-less accurate for expensive homes. Removing observations 169 and 1472 improved model fit marginally (R^2 +0.72%, RSE -3.3%) but was justified by more than statistics alone as these homes had an abnormal sale condition with extraordinarily low sale prices ($12,789 and $13,100).

These analytical activities do create difficulties such as increased documentation burden, interpretation complexity, and the ever-present risk of researcher degrees of freedom enabling p-hacking; but when conducted transparently with pre-established criteria, theoretical justification, and sensitivity analysis, they transform an invalid model into a trustworthy one. The benefits of valid inference and improved accuracy far outweigh the complexities introduced.

Statistical hypothesis tests in regression can be trusted when assumptions are met, but with caution about what they are telling us. Model 6's hypothesis tests are trustworthy because diagnostic plots confirm excellent assumption compliance: constant variance across all predicted values, near-perfect normality of residuals, no multicollinearity (VIF <2.5). The overwhelming evidence (all p<2e-16, t-values of 11-40) makes it implausible that random chance explains these relationships. However, Model 6 likely omitted other important variables for predicting SalePrice, meaning coefficients may partly reflect confounding with unobserved factors.

Several important next steps would improve and validate this model. Most immediately, the extreme outlier in Model 6 (standardized residual ≈ -11) requires investigation to determine if it's another abnormal sale requiring removal or a legitimate unusual property revealing model limitations. Adding categorical variables, particularly Neighborhood, represents the largest untapped opportunity, and potentially increasing R^2 value since location drives enormous price variation through school quality, amenities, and prestige.

A. Appendix

Figure A1. Diagnostic plots for Model 1

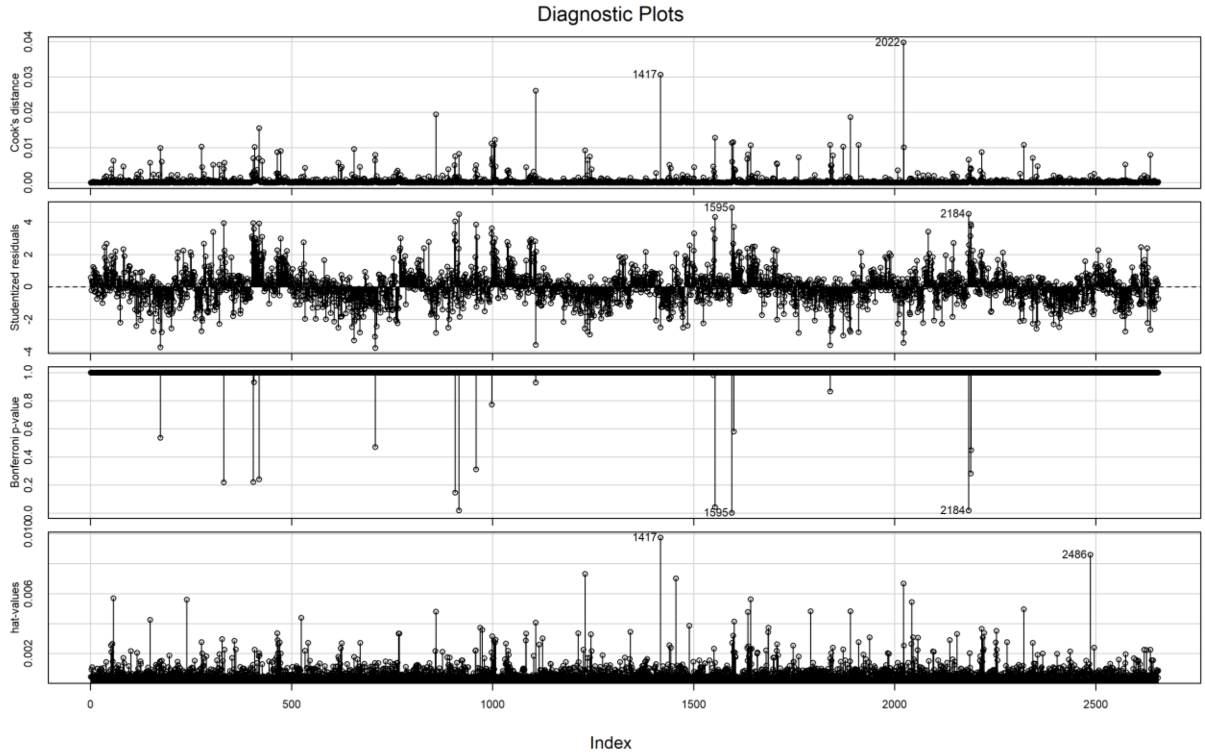


Figure A2. Bubble plot of Influential Observations in Model 1

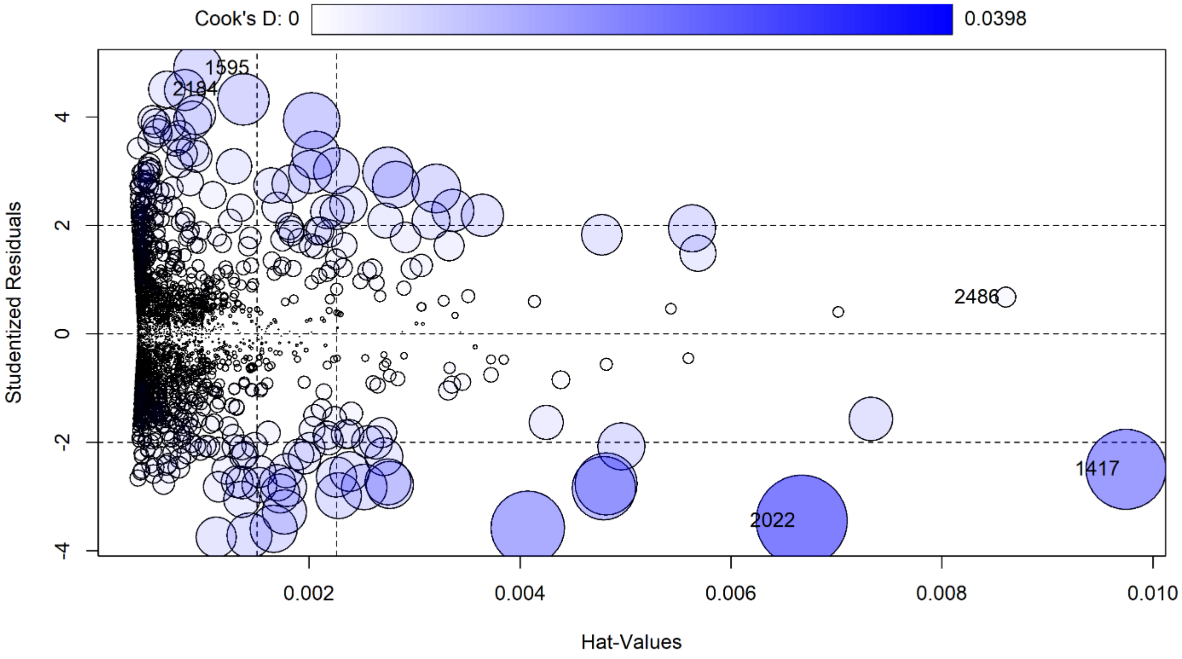


Figure A3. Diagnostic Plots for Model 2

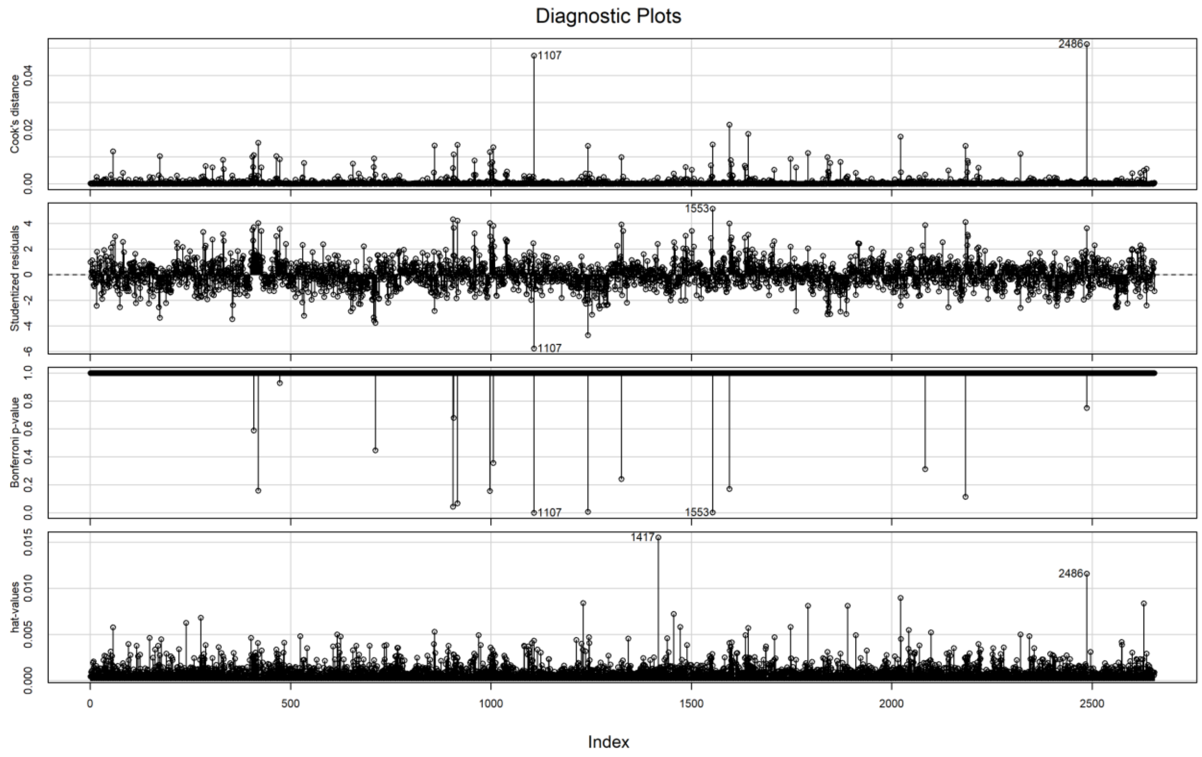


Figure A4. Bubble Plot of Influential Observations in Model 2

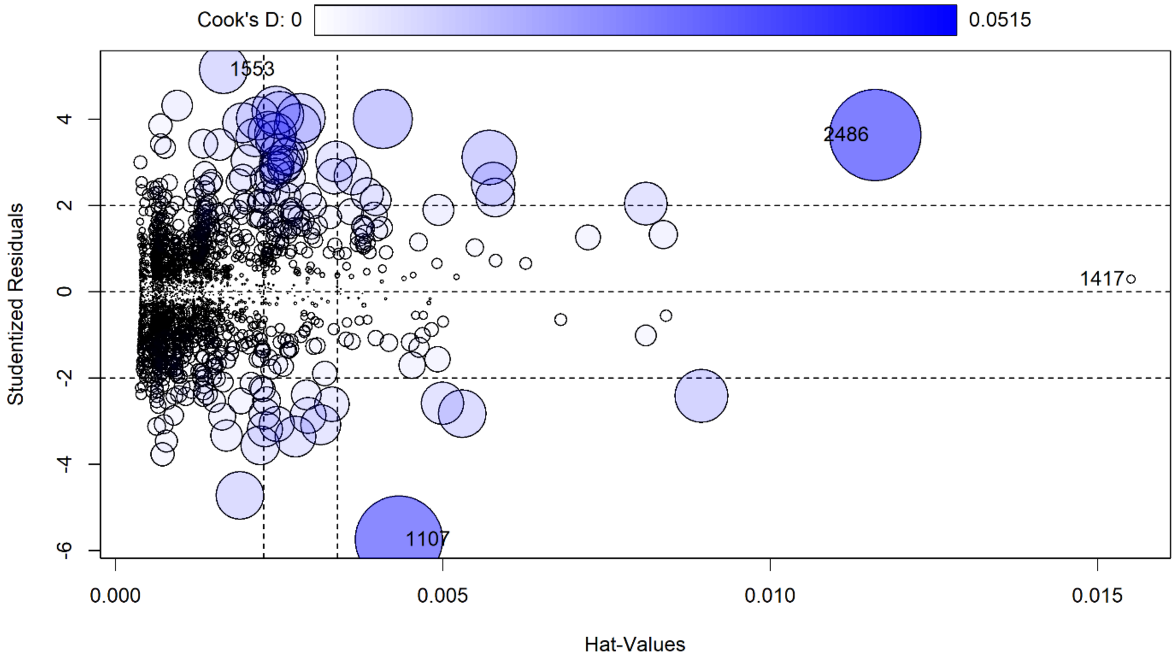


Figure A5. Diagnostic Plots for Model 3

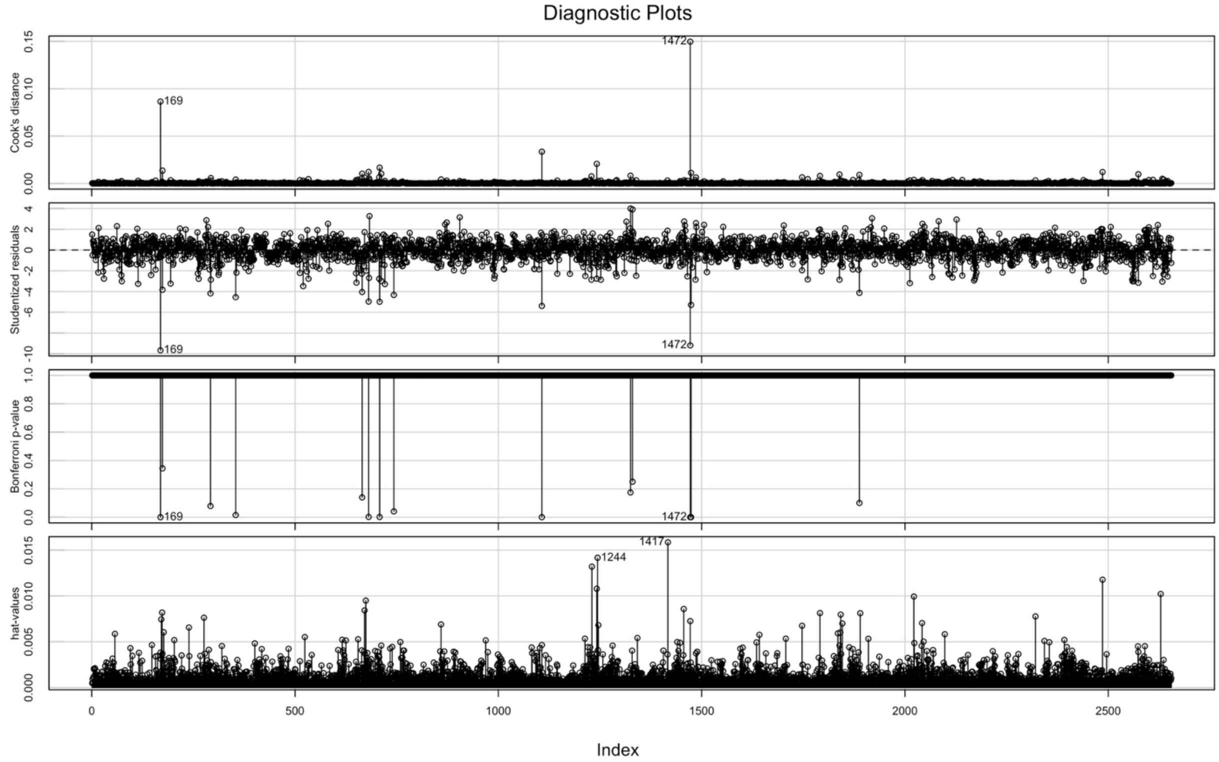
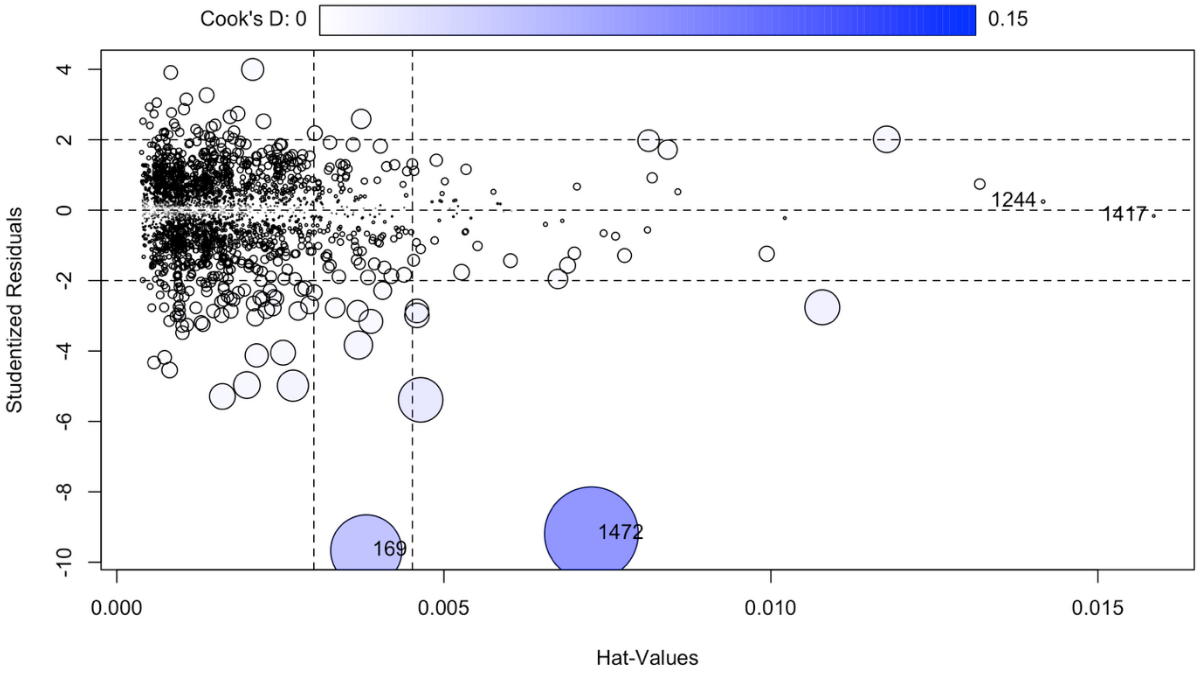


Figure A6. Bubble Plot of Influential Observations in Model 3


Cook's D: 0
Studentized Residuals
Hat-Values