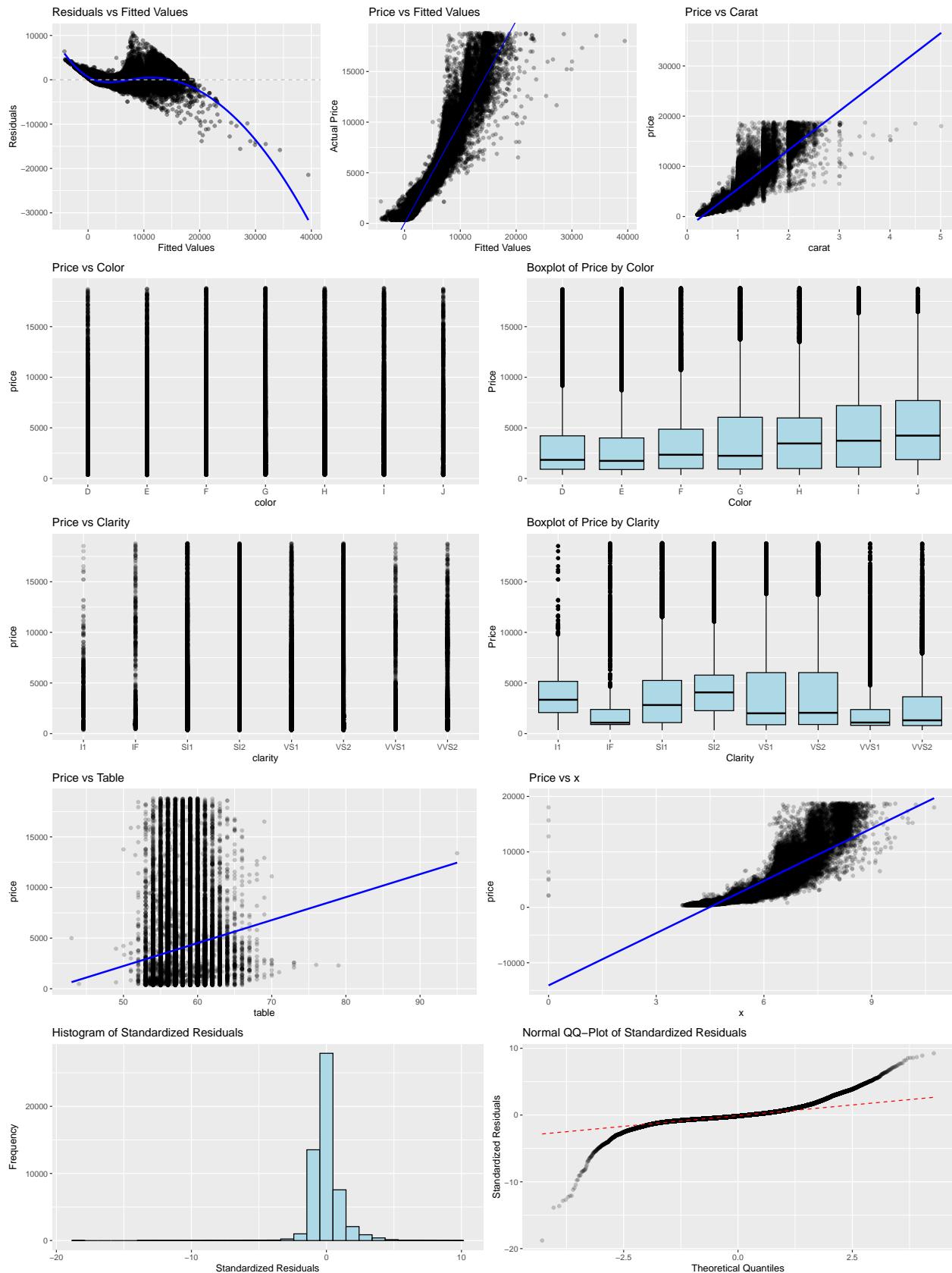


Introduction

Data Description

summarize numerically or graphically (in a single figure/table) each predictor in your dataset that will be used in the preliminary model, and interpret the descriptive statistics in the context of what the predictors measure and how it relates to the research question



Preliminary Model Results

$$\begin{aligned}
y = & \beta_0 + \beta_1 \cdot \text{carat} + \beta_2 \cdot I(\text{color} = E) + \beta_3 \cdot I(\text{color} = F) + \beta_4 \cdot I(\text{color} = G) + \beta_5 \cdot I(\text{color} = H) \\
& + \beta_6 \cdot I(\text{color} = I) + \beta_7 \cdot I(\text{color} = J) + \beta_8 \cdot I(\text{clarity} = \text{IF}) + \beta_9 \cdot I(\text{clarity} = \text{SI1}) \\
& + \beta_{10} \cdot I(\text{clarity} = \text{SI2}) + \beta_{11} \cdot I(\text{clarity} = \text{VS1}) + \beta_{12} \cdot I(\text{clarity} = \text{VS2}) \\
& + \beta_{13} \cdot I(\text{clarity} = \text{VVS1}) + \beta_{14} \cdot I(\text{clarity} = \text{VVS2}) + \beta_{15} \cdot \text{table} + \beta_{16} \cdot x + \epsilon
\end{aligned}$$

Assume $\mathbb{E}[\epsilon] = 0$, $\mathbb{V}[\epsilon] = \sigma^2$, and $\epsilon \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

$$\begin{aligned}
\hat{E}[y] = \hat{y} = & \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{carat} + \hat{\beta}_2 \cdot I(\text{color} = E) + \hat{\beta}_3 \cdot I(\text{color} = F) + \hat{\beta}_4 \cdot I(\text{color} = G) + \hat{\beta}_5 \cdot I(\text{color} = H) \\
& + \hat{\beta}_6 \cdot I(\text{color} = I) + \hat{\beta}_7 \cdot I(\text{color} = J) + \hat{\beta}_8 \cdot I(\text{clarity} = \text{IF}) + \hat{\beta}_9 \cdot I(\text{clarity} = \text{SI1}) \\
& + \hat{\beta}_{10} \cdot I(\text{clarity} = \text{SI2}) + \hat{\beta}_{11} \cdot I(\text{clarity} = \text{VS1}) + \hat{\beta}_{12} \cdot I(\text{clarity} = \text{VS2}) \\
& + \hat{\beta}_{13} \cdot I(\text{clarity} = \text{VVS1}) + \hat{\beta}_{14} \cdot I(\text{clarity} = \text{VVS2}) + \hat{\beta}_{15} \cdot \text{table} + \hat{\beta}_{16} \cdot x
\end{aligned}$$

Table 1: Numerical Summary of Coefficient Estimates

Term	Estimate
(Intercept)	-1388.839
carat	10945.695
colorE	-210.489
colorF	-286.752
colorG	-493.858
colorH	-993.809
colorI	-1474.570
colorJ	-2387.941
clarityIF	5665.892
claritySI1	3894.266
claritySI2	2926.092
clarityVS1	4845.059
clarityVS2	4528.148
clarityVVS1	5309.983
clarityVVS2	5234.228
table	-32.708
x	-900.270

The residual analysis revealed significant violations of core regression model assumptions: linearity, constant error variance (homoscedasticity), and normality of errors. The **Residuals vs Fitted Values** plot showed a clear, curved LOESS trendline with increasing spread as fitted values increased, suggesting non-linearity and heteroscedasticity.

Similarly, the **Price vs Fitted Values** plot showed a generally strong positive relationship, but with slight curvature and growing variance at higher price levels, this causes concern about unequal error variance—indicating potential non-linearity and heteroscedasticity. The **Price vs Carat** scatterplot suggested an exponential relationship that a smile linear model fails to capture reasonably.

At first glance, categorical variables **color** and **clarity** don't appear to be strong predictors of price due to seemingly constant spread of points across levels. However, the boxplots reveal significant stratified effects on **price**, confirming that these variables do show variation.

For **color**, although the maximum prices are similar across all levels, the median prices differ, with diamonds in **color J** having higher median prices, and D showing more high outliers.

Diamonds with **clarity** I1 show outliers and higher median prices, whereas those with IF appear to have more outliers and lower medians. This stratification reinforces the idea that both are valid categorical predictors for diamond price modeling.

However, the **Price vs Table** plot showed a weak and noisy association, with data points highly concentrated in a narrow range, suggesting that the table predictor may not be a strong predictor to predict price. The **Price vs x** plot also showed curvature, suggesting non-linearity.

Furthermore, the **histogram of standardized residuals** sharply peaked around 0 with heavy tails, and the **normal QQ-plot** confirmed it as it showed deviations at both tails, indicating violations of the normality assumption.

From model estimates, a diamond with **color** D, **clarity** I1, and 0 values for **table**, **x**, and **carat**, is predicted to have an average price of -\$1,388.389. Though this scenario isn't practically meaningful, it serves as the model's (baseline) reference level. When the diamond's **color** is E and **clarity** is IF, with the same 0 values for the continuous variables, the (estimated) average price increases to \$4,066.564, reflecting the strong positive contribution of high **clarity** IF to the overall **price**.

Moreover, a 1-gram increase in **carat**—with the same values as the previous example—raises the predicted average price by \$10,945.695 per gram of **carat**, confirming that **carat** is the most influential predictor.

Interestingly, **table** and **x** are both negatively associated with **price**, with coefficients of -\$32.71 and -\$900.27, respectively. This is counterintuitive as larger dimensions are generally more desirable. **x** likely reflects multicollinearity with **carat**, since both capture size-related information. This redundancy suggests the need to reconsider variable inclusion—possibility removing **x** or replacing with a composite variable like volume.

Bibliography