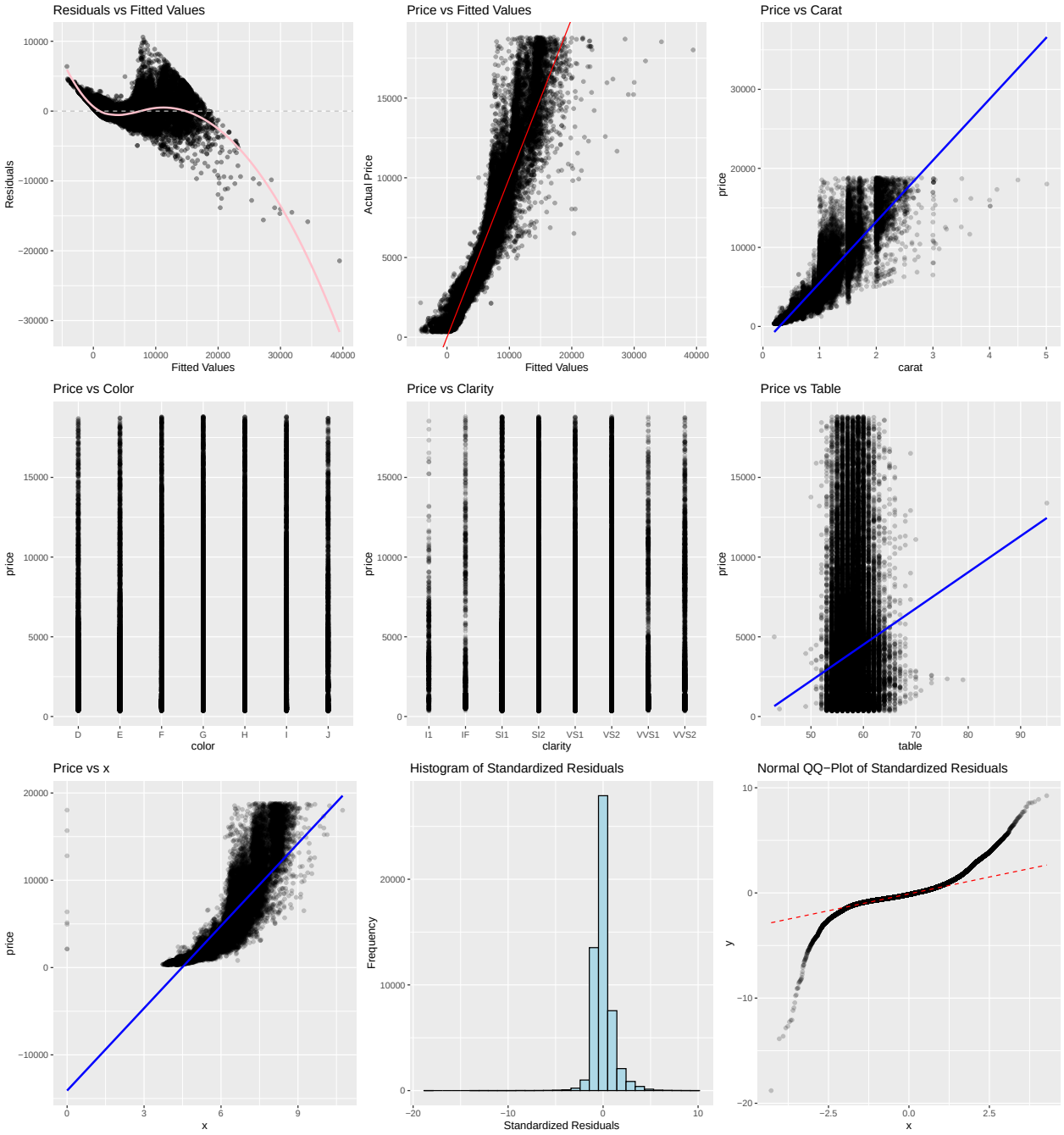
Introduction

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 descrip8ve sta8s8cs in the context of what the predictors measure and how it relates to the research ques8on



Preliminary Model Results

To investigate how the carat—price relationship varies across different diamond clarity levels after adjusting for color, table size, and x-dimension (length), we fit a Multiple Linear Regression model: **price** ~ **carat** +

color + clarity + table + x.

The residual analysis revealed notable violations of key 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 revealed slight curvature with growing variance at higher price levels, causing concerns 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. Predictors such as color and clarity showed stratified effects on price, validating them as categorical predictors. 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.

Despite these concerns, the model explains a substantial portion of the variance in price, with an adjusted R^2=0.9174.

The coefficient estimates revealed that **carat** is the strongest numerical predictor: for every additional **carat**, **price** increases by approximately $10,945.695. Among the categorical predictors, diamonds graded as clarityIF (Internally Flawless) are associated with a $5,665.892 higher average price than the base-level (I1), holding other variables constant. Lower **color** grades such as J result in significant price reductions, approximately $2,387.941 less than D.

Interestingly, numerical predictors **table** and **x** showed negative associations with **price**. Holding other variables constant, a 1-unit increase in the **table** value is associated with an estimated $32.708 decrease in **price**. This result is counterintuitive, as diamonds with slightly larger **table** sizes are generally more desirable for their ability to reflect light effectively, so increasing their value. Similarly, a 1-unit increase in the x-dimension corresponds to an estimated $900.27 decrease in **price**, which is also unexpected since longer diamonds are usually more valuable. However, because **x** is strongly correlated with **carat**, this suggests potential multicollinearity—both variables may be capturing similar variance in the data. The large negative coefficient for **x** likely reflects this redundancy. So one might consider centering or removing **x** or replacing it with a composite variable like volume.

These findings suggest that while the model identifies important trends and supports the hypothesis that clarity influences the carat-price relationship, assumption violations and multicollinearity call attention for further model refinement.

Bibliography