Dazzling Predictions: Unveiling a Model for Diamond Price Projection

Mathematical Statistics - Research Project

Filippo Antonio Ronzino, Giuseppe Iannone

January 2024

1 Introduction

In our research we will explore the diamond market, focusing specifically on understanding the relationship between the characteristics of a diamond and its price. With the exception of highly unique diamonds typically sold at auctions with distinctive prices, we expect a robust correlation between diamond features and their prices. Our ultimate goal is to construct a predictive model capable of estimating the price of a diamond based on its characteristics, using a multivariate linear regression model. This analysis aims to provide a clear perspective on the relative value of diamonds and could have significant implications for the market, facilitating informed decisions for buyers, sellers, and industry operators.

2 Dataset

2.1 Sources

For our analysis we use a dataset [1] that contains the prices and other attributes of almost 54,000 diamonds.
A schematic description of such attributes, with the abbreviations used in the R script, follows:

- `carat` weight of the diamond
- `cut` quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- `color` diamond colour, from J (worst) to D (best)
- `clarity` a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- `depth` total depth percentage = $\frac{z}{mean(x,y)}$
- `table` width of top of diamond relative to widest point
- `price` price in US dollars
- `x` length in mm
- `y` width in mm
- `z` depth in mm

2.2 Data Cleaning & Data Visualization

Firstly we proceed with the data cleaning. Calling summary of our data set in R, it is clear that possible error of measurement were registered. An instance of this is that the columns `x`, `y` and `z` have a minimum value that is zero, which reasonably represent a discrepancy in collecting the data.

Moreover, there are three categorical variables, i.e. "clarity", "cut", "color", that have to be converted in $k-1$ dummy variables, with $k :=$ # of categories for each factor. Furthermore, we make sure that there are no NAs in our dataset.

Finally, we deal with outliers that are easily found by boxplotting the corresponding data for each category (in Figure 1 outliers for each category are shown in red).

[1] Link to the dataset: https://www.kaggle.com/datasets/shivam2503/diamonds

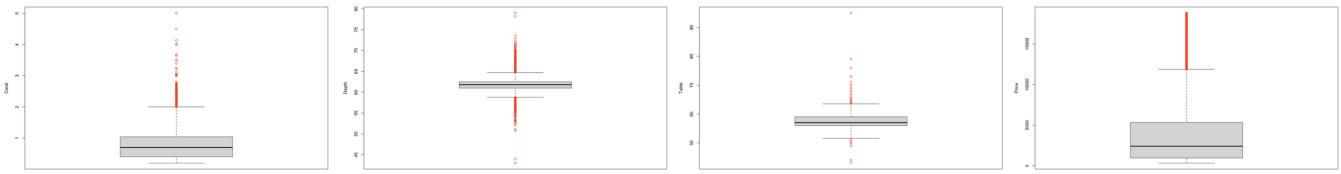


Figure 1: from left to right `carat`, `depth`, `table`, `price`

Now, once the outliers are removed from the dataset, we proceed with creating the correlation matrix and looking for possible dependent covariates that have covariance close to 1. Indeed, we find, as Figure 2 shows, mutual dependence between `x`, `y`, `z` (which is explained by the fact that they represent a volume quantity) and moreover an overall dependence with `carat`; consequently, we remove the columns `x`, `y`, `z` and just keep the `carat` one.

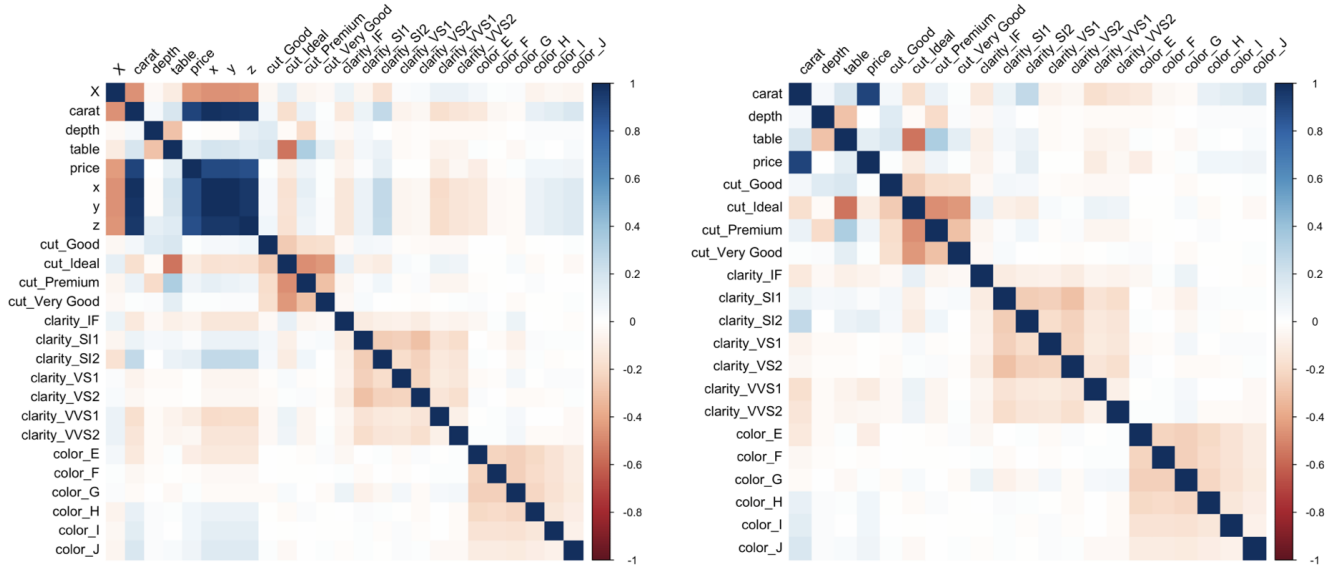


Figure 2: Correlation matrix before and after the removal of `x`, `y`, `z`

Having a clear dataset, we are almost ready for the multilinear regression. First we check which kind of relation exists between the numerical covariates and the price of the diamond; therefore we plot the graph `carat`-`price` (noticing a squared correlation that will induce a `carat` squared component in the model, as shown in Figure 3), while we do not notice any particular pattern in the other cases.

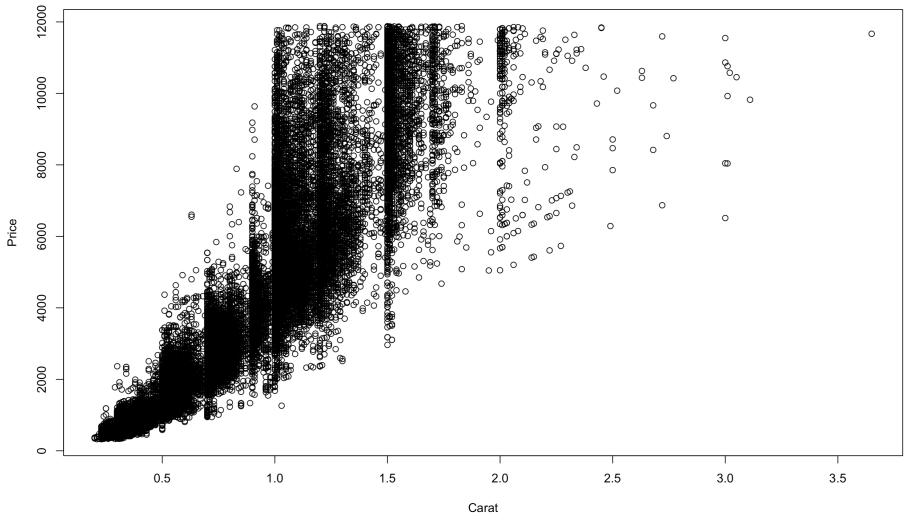


Figure 3: `carat` - `price`

3 Multivariate Linear Regression

We create our linear model, including all the covariates and in particular the squared `carat` term. Summarizing, we obtain a good model, with $R^2 = 0.9172$ close to 1, and that as a whole is statistically significant, since F-statistic is $2.657 \cdot 10^4$, and the associated p-value is very below $\alpha := 0.05$. Moreover the p-values associated with the t-statistics for each coefficient are below α, so we have enough in evidence to say that the covariates together are useful for our model.

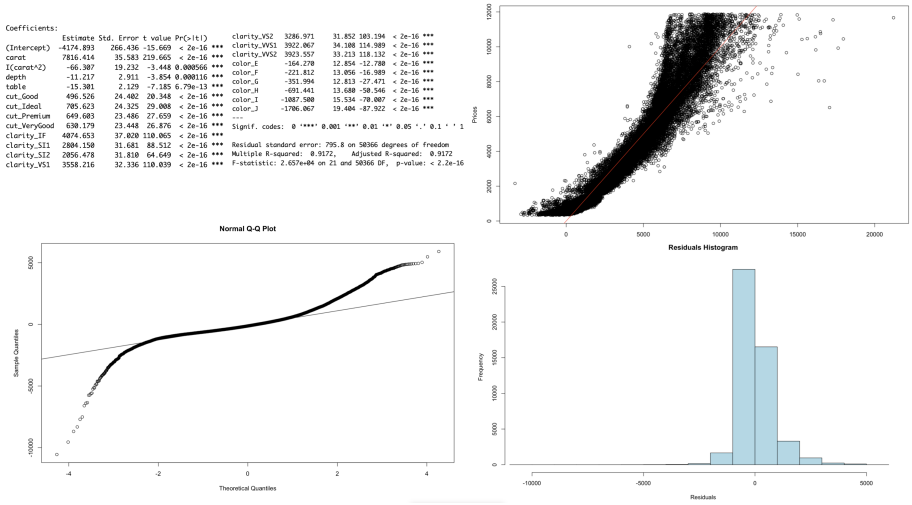


Figure 4: Summary, Predicted vs actual prices, Q-Q Plot and Histogram of residuals

Performing now the model validation, i.e. a regression diagnostic, it is evident that the distribution of the residuals is not properly normal [2], having heavier tails than that of a normal distribution and indeed the histogram reveals the asymmetry of it. Furthermore, we clearly see that the scatter plot in Figure 5 shows a violation of homoscedasticity assumption of the noise term, that is $e_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, ..., n$. Consequently, we update our model in order to correct the latter with a Cox-Box transformation of the type $t(Y) = Y^{1/3}$. As expected, this modification takes the adjusted $R^2 = 0.9623$ which has significantly increased from the previous attempt and moreover produce a more aligned scatter plot of predicted against the actual prices, and the variance of the residuals is now closer to be constant, since it does not increase as much as it did before (Figure 5).

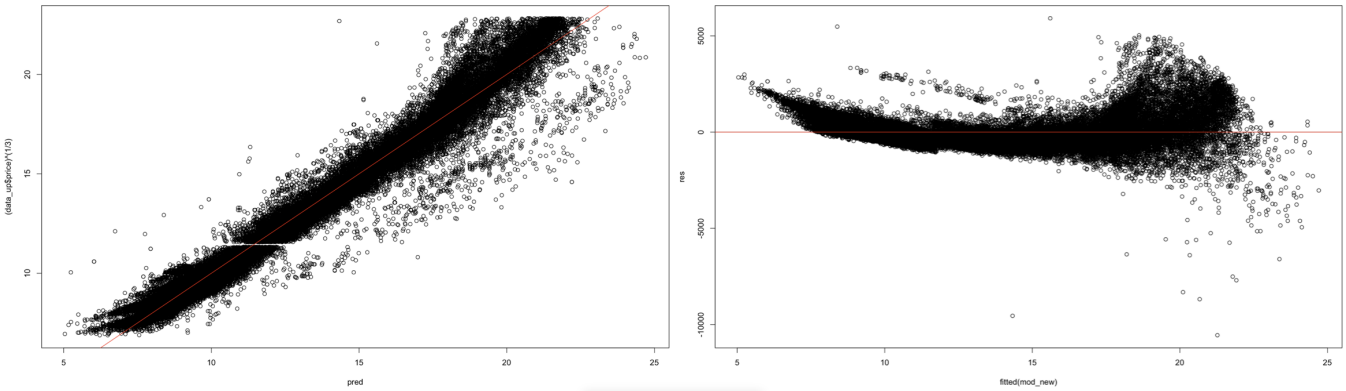


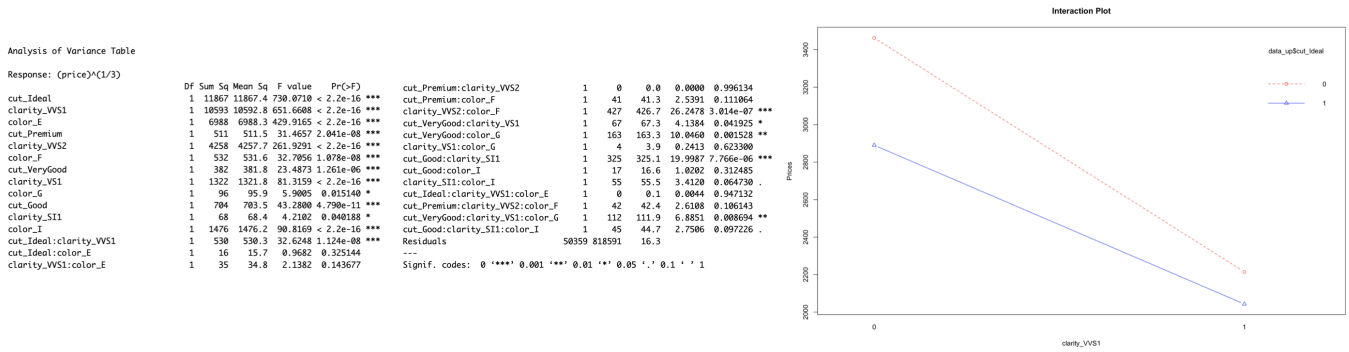
Figure 5: Summary, Predicted vs actual prices, Q-Q Plot and Histogram of residuals

3.1 ANOVA

The presence of categorical variables suggests us to check for the existence of interaction effects between our factors. To look for those, we perform the ANOVA. Before diving into it, we make an interaction plot between `clarity_VVS1`

[2] We refer to section 4 for the Kolmogorov-Smirnov test that will confirm that a statistically significant deviation from normality exists in the residuals of the final updated model.

and `cut_Ideal` to have a qualitative understanding of the connection of the two: basically we are supposing that there is an interaction effect on the price of the diamond if it has a particular good (or bad) combinations of attributes. As the Figure 6 shows, it's reasonable to check if this interaction effect is statistically significant, so we perform the ANOVA, using the good (and bad) combinations of `color`, `cut` and `clarity`
In the ANOVA table (Figure 6), we can see that more than one interaction is statistically significant, so we decide to put the interactions of these factors in a new updated linear model



```
Analysis of Variance Table

Response: (price)^(1/3)
                          Df Sum Sq Mean Sq  F value    Pr(>F)
cut_Ideal                  1  11867 11867.4 730.0710 < 2.2e-16 ***
clarity_VVS1               1  10593 10592.8 651.6608 < 2.2e-16 ***
color_E                    1   6988  6988.3 429.9165 < 2.2e-16 ***
cut_Premium                1    511   511.5  31.4657 2.041e-08 ***
clarity_VVS2               1   4258  4257.7 261.9291 < 2.2e-16 ***
color_F                    1    532   531.6  32.7056 1.078e-08 ***
cut_VeryGood               1    382   381.8  23.4873 1.261e-06 ***
clarity_VS1                1   1322  1321.8  81.3159 < 2.2e-16 ***
color_G                    1     96    95.9   5.9005  0.015140 *
cut_Good                   1    704   703.5  43.2800 4.790e-11 ***
clarity_SI1                1     68    68.4   4.2102  0.040188 *
color_I                    1   1476  1476.2  90.8169 < 2.2e-16 ***
cut_Ideal:clarity_VVS1     1    530   530.3  32.6248 1.124e-08 ***
cut_Ideal:color_E          1     16    15.7   0.9682  0.325144
clarity_VVS1:color_E       1     35    34.8   2.1382  0.143677
cut_Premium:clarity_VVS2   1      0     0.0   0.0000  0.996134
cut_Premium:color_F        1     41    41.3   2.5391  0.111064
clarity_VVS2:color_F       1    427   426.7  26.2478 3.014e-07 ***
cut_VeryGood:clarity_VS1   1     67    67.3   4.1384  0.041925 *
cut_VeryGood:color_G       1    163   163.3  10.0460  0.001528 **
clarity_VS1:color_G        1      4     3.9   0.2413  0.623300
cut_Good:clarity_SI1       1    325   325.1  19.9987 7.766e-06 ***
cut_Good:color_I           1     17    16.6   1.0202  0.312485
clarity_SI1:color_I        1     55    55.5   3.4120  0.064730 .
cut_Ideal:clarity_VVS1:color_E  1      0     0.1   0.0044  0.947132
cut_Premium:clarity_VVS2:color_F  1     42    42.4   2.6108  0.106143
cut_VeryGood:clarity_VS1:color_G  1    112   111.9   6.8851  0.008694 **
cut_Good:clarity_SI1:color_I  1     45    44.7   2.7506  0.097226 .
Residuals                 50359 818591    16.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Figure 6: ANOVA table and Interaction `cut_Ideal`, `clarity_VVS1`

3.2 Step Down

In our first linear model, all the covariates were statistically significant, while in our new updated one the p-values associated with the t-statistics (for testing $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0 \; \forall i = 1, .., n$) for some coefficient is $> \alpha$. To streamline the model, we adopt a stepwise approach, i.e. sistematically start removing the covariate with the highest associated p-value, examine the new linear model, and continue to remove the covariate with the highest p-value, until all the covariates have an associated p-value $\leq \alpha$. This procedure is useful because it allows us to simplify the model, without losing statistically significant information given by the covariates.

3.3 BIC

In order to choose between the smaller model and the bigger ones an alternative would be to use information criteria. In our case, due to the size of the dataset, we choose the BIC: this criterion penalizes the bigger models (so higher number of covariates) but also takes into account the number of observations. We run BIC on the four models obtained in the previous step down and, against expectations, we notice that the larger model has the lowest BIC value. Hence, while step down suggests to prefer the smaller model, BIC, instead, advises for the larger one.

4 Results & Testing

Analysing the results of the final model, i.e. the smaller one, we get an adjusted $R^2 = 0.9757$, with a terrific improvement, and that as a whole is statistically significant, since F-statistic is $8.417 \cdot 10^4$, and the associated p-value is very below $\alpha := 0.05$. Moreover the p-values associated with the t-statistics for each coefficient are below α, so we have enough evidence to say that the covariates together are useful for our model. The plot of predicted prices against actual prices shows a line that fits very well the data, and the distribution of the residuals, although non-Normal, shows some improvement from the one of the first model.

```
Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)         0.700222   0.215735    3.246 0.001172 **
carat              18.416941   0.028811  639.228  < 2e-16 ***
I(carat^2)         -3.879367   0.015570 -249.153  < 2e-16 ***
depth              -0.009720   0.002357   -4.125 3.72e-05 ***
table              -0.011438   0.001724   -6.634 3.31e-11 ***
cut_Good            0.437497   0.019758   22.143  < 2e-16 ***
cut_Ideal           0.773923   0.019745   39.195  < 2e-16 ***
cut_Premium         0.654103   0.019018   34.394  < 2e-16 ***
cut_VeryGood        0.595814   0.019178   31.068  < 2e-16 ***
clarity_IF          4.909594   0.029977  163.781  < 2e-16 ***
clarity_SI1         2.899753   0.025651  113.047  < 2e-16 ***
clarity_SI2         2.020988   0.025754   78.472  < 2e-16 ***
clarity_VS1         3.911680   0.026482  147.709  < 2e-16 ***
clarity_VS2         3.557959   0.025788  137.968  < 2e-16 ***
clarity_VVS2        4.374065   0.027331  160.041  < 2e-16 ***
color_E            -0.251380   0.010407  -24.155  < 2e-16 ***
color_F            -0.418458   0.010853  -38.556  < 2e-16 ***
color_G            -0.652605   0.010377  -62.888  < 2e-16 ***
color_H            -1.105493   0.011075  -99.820  < 2e-16 ***
color_I            -1.664694   0.012577 -132.361  < 2e-16 ***
color_J            -2.354212   0.015710 -149.856  < 2e-16 ***
clarity_VVS1        4.651672   0.030274  153.650  < 2e-16 ***
cut_Ideal:clarity_VVS1  -0.110126   0.022783   -4.834 1.34e-06 ***
clarity_VVS2:color_F     0.187225   0.024690    7.583 3.43e-14 ***
cut_VeryGood:clarity_VS1 -0.066153   0.019347   -3.419 0.000628 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7: Summary

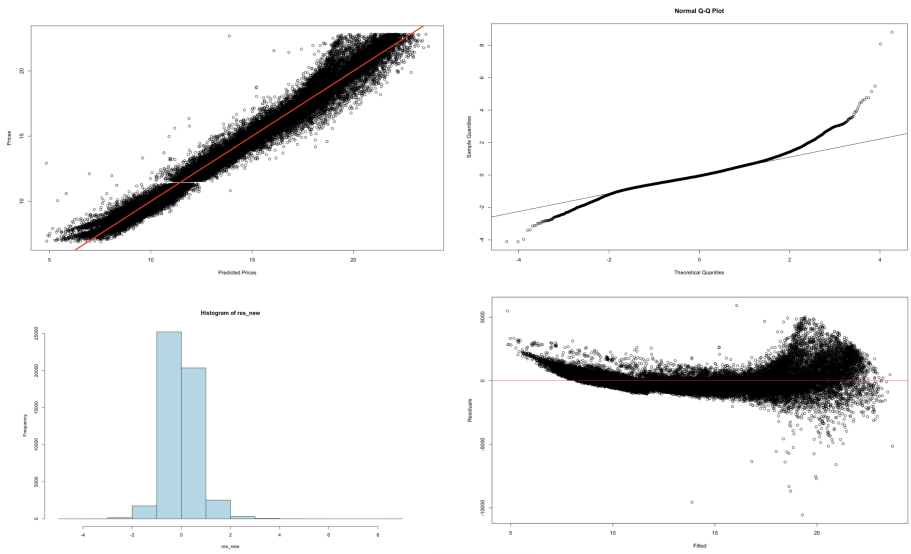


Figure 8: Predicted-Actual prices, QQ-Plot of residuals, Histograms of residuals, Variance of residuals

Moreover the p-values associated with the t-statistics for each coefficient are below α, so we have enough evidence to say that the covariates, and the interactions between some of them, all together are useful for our model. The plot of predicted prices against actual prices evidences a strong alignment, and the distribution of the residuals, although clearly non-Normal (by the QQ-Plot that shows some heavy tails, the histograms and the Fitted against Residuals) shows some improvement from the one of the first model (Figure 5).

To confirm the non-normality of the residuals, we run a Kolmogorov-Smirnov test. So we create a random sample from a Normal distribution with $\mu := \bar{X}$ and $\sigma^2 = S_X^2$, and we look for $D := \sup_x |\mathbb{F}_n(x) - F_{\mu,\sigma^2}(x)|$; if D sufficiently high, we make the strong conclusion that the residuals are not normal, and indeed we get $D = 0.15853$ whose associated p-value $< 2.2 \cdot 10^{-16}$, hence we have statistically significant evidence that the residuals are not normally distributed.

5 Conclusion

In concluding our work, we will provide some closing observations. Most of the updates in the model and corrections made have been discussed along the way as they were introduced and hence we will briefly mention them summarizing our path.

We highlight the significance of our research in constructing a predictive model for diamond prices, using a multivariate linear regression framework. Throughout our methodological process, we addressed data anomalies, specifically outliers and covariate dependencies, ensuring the robustness of our analyses. The resultant model, characterized by a satisfactory coefficient of determination and statistical significance, incorporated essential diamond attributes. Notably, the persistence of non-normal residuals in our regression diagnostics prompted us to employ a Cox-Box transformation to mitigate this issue. Furthermore, in refining our model, we undertook a systematic step-down approach using ANOVA to examine potential interaction effects between the categorical variables. The identified statistically significant interactions were then integrated into an updated and final linear model. Instead, using the BIC we should prefer the larger model that takes into account all interaction effects. Due to this contradiction, it would be useful to use other methods for model selection, for example LASSO that penalizes the model according to the parameter λ. Despite the model's strengths, the presence of non-normal residuals - clearly quantitatively explained by Kolmogorov-Smirnov test - and still not perfect homoscedasticity, signals an avenue for further exploration and refinement, emphasizing the complex interplay between diamond features and market values that might not be captured by a linear model. Another possible improvement in the prediction of prices could be a k-fold cross-validation, i.e. splitting the data set into k parts of equal size and using each part once for validation while the union of the rest for training; this model could be better at training and worse at testing.

The final model, although with some limitations, demonstrates the potential of linear regression in predicting diamond prices. However, further refinement and exploration of other modeling techniques are warranted to fully capture the complexity of diamond prices.