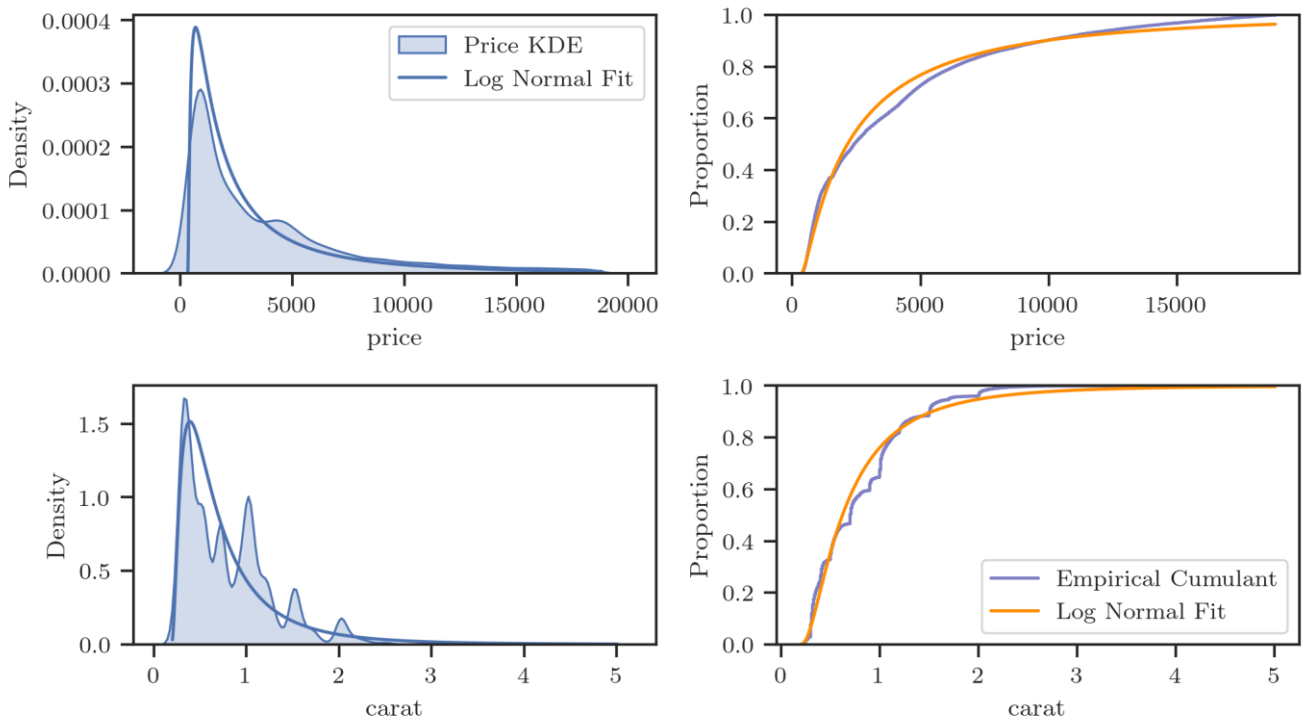
# PREDICTION OF DIAMOND PRICE FROM CARAT

*Elisa Omodei, Leonardo Di Gaetano*
Central European University
e-mails: di-gaetano_leonardo@phd.ceu.edu

## 1 INTRODUCTION

This report advances our exploration of the Seaborn 'diamonds' dataset, which was introduced in our earlier work. Here, we focus on the correlation between a diamond's carat weight and its price. By analyzing this relationship, our objective is to ultimately develop a predictive model capable of estimating a diamond's price from its carats. This step is crucial for anyone looking to understand or predict diamond prices based on their size, blending data analysis with practical application.

## 2 DISTRIBUTION AND FITTING




In our pursuit to understand the relationship between diamond carat weight and price, we began with an examination of their distributions in the dataset. Identifying an accurate distribution model for these variables is crucial for our correlation study. To this end, we tested both the normal and log-normal distributions to see which provided the best fit.

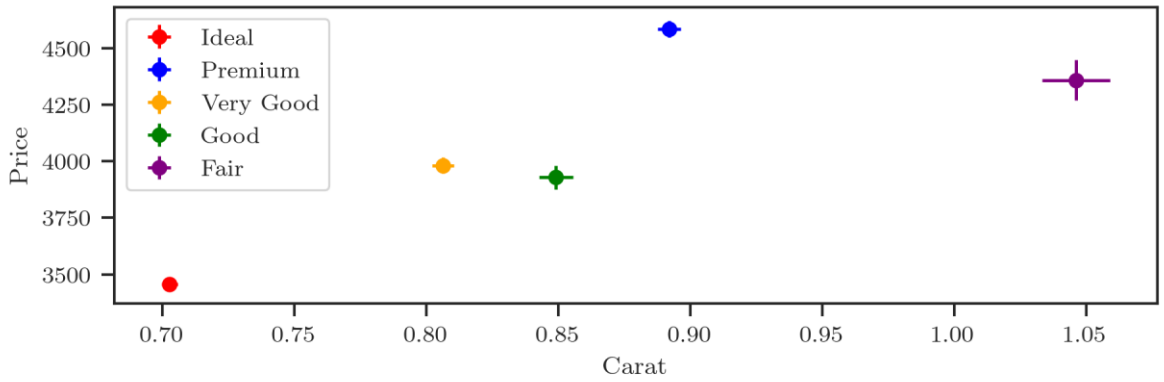
Our analysis, detailed in the accompanying notebook, involved statistical tests to determine if we could dismiss either distribution as a potential model for our data. These tests did not lead us to reject either the normal or log-normal distributions outright. However, further examination through Q-Q plot comparisons showed that the log-normal distribution more closely matches the observed data for both variables.

Given these findings, we selected the log-normal distribution for our further analysis. This choice is supported by visual evidence in the Q-Q plots presented within the report. We encourage reviewing the notebook for a detailed comparison and the rationale behind not choosing the normal distribution for our analysis.

## 3. DATA CLASSIFICATION

As outlined in our initial report, the diamonds dataset categorizes diamonds into five quality classes: Ideal, Premium, Very Good, Good, and Fair. In this section, we explore the feasibility of distinguishing these classes based on our analysis. This investigation aims to understand if the dataset provides a clear basis for differentiating diamonds by quality, thereby adding another dimension to our study of diamond pricing and characteristics**.**

### 3.1 ERROR BARS AND CONFIDENCE INTERVAL




In our initial step to differentiate between the diamond quality classes defined in the dataset (Ideal, Premium, Very Good, Good, Fair), we conducted a comparative analysis focusing on the mean price and mean carat for each class. To enhance this comparison, we incorporated error bars into our plots. These error bars represent the standard error of the mean (SEM), calculated as the standard deviation divided by the square root of the sample size for each quality class. This statistical measure provides an estimate of the uncertainty associated with the mean values of price and carat for each class. Observing the figure above, it becomes evident that the five classes can be visually distinguished based on these criteria, illustrating the variability and distinctiveness of each class in terms of mean price and carat size.

### 3.2 HYPOTHESIS TESTING

To further substantiate our observation that the five diamond quality classes (Ideal, Premium, Very Good, Good, Fair) can be visually distinguished based on mean price and carat, we employed two statistical tests. Our primary goal was to validate whether the classes indeed differ statistically in terms of their pricing and carat distribution.

Initially, we applied the Analysis of Variance (ANOVA) test to determine if there's a significant difference in either price or carat weight across the classes. The ANOVA test yielded a positive outcome, indicating that at least one class is statistically distinguishable from the others, with a p-value less than 5%. This result, detailed in the accompanying notebook, confirmed that the differences observed in the mean values and error bars were not merely visual but also statistically significant.

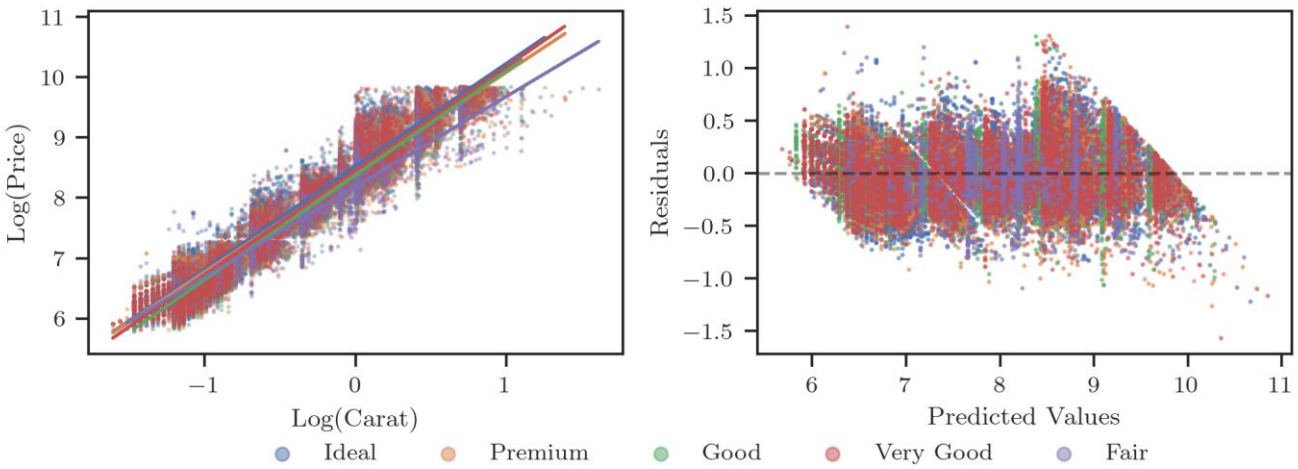
Following this, we verified that the standard deviations within each class were comparable—this analysis is also documented in the notebook—to ensure the applicability of further pairwise comparisons. Satisfied with this condition, we proceeded to conduct t-tests between every pair of quality classes. These t-tests aimed to identify specific differences between each pair of classes, providing a more granular view of the distinctions among them. The outcomes of these pairwise t-tests are summarized in the table below. This table presents a detailed view of the statistical significance between each pair of diamond quality classes, highlighting those comparisons where the differences in mean price and carat are not only apparent but also statistically substantiated. According to what is shown in the previous figure, almost all the diamond class can be distinguished. In particular, except for the comparison between the 'Good' and 'Very Good' classes, all other pairs showed statistically significant differences through this test, confirming the distinctiveness of the quality classes as suggested by the earlier figure.

| | | p-value |
|---|---|---|
| Ideal | Premium | 0 |
| Ideal | Good | 0 |
| Ideal | Very Good | 0 |
| Ideal | Fair | 0 |
| Premium | Good | 0 |
| Premium | Very Good | 0 |
| Premium | Fair | 0.0191 |
| Good | Very Good | 0.4056 |
| Good | Fair | 0 |
| Very Good | Fair | 0.0001 |

## 4. REGRESSION AND PREDICTION

### 4.1 LINEAR REGRESSION AND RESIDUALS

In this final section, we turn our attention to the regression analysis between the carat weight and price of diamonds. Building on the foundation that diamond quality classes can indeed be distinguished, as evidenced by our previous analyses, we proceed to conduct separate regression analyses for each of the five quality classes. This approach allows us to tailor our predictive model to each class, recognizing that the relationship between carat weight and price may vary significantly across different quality categories. By conducting five distinct regressions, one for each class, we aim to capture these nuances and provide more accurate predictions for a diamond's price based on its carat weight within each quality classification.




Following a visual assessment of the scatter plots between carat weight and diamond price (see the notebook) it becomes apparent that the points exhibit a linear correlation when plotted on a log-log scale. This observation prompted us to pursue a linear regression analysis using the logarithmic values of both price and carat. The outcomes of this approach are illustrated in the first panel of the figure presented above, where we display the linear fit between the log-transformed variables.

The second panel of the figure offers an examination of the residuals from this regression model. These residuals, the differences between the observed and predicted values, are distributed homogeneously above and below the regression line, indicating a well-fitting model. Notably, this log-transformed linear regression demonstrates superior performance over a linear regression conducted on the original, non-transformed data. For a comparative analysis of the linear fit on the original dataset, we refer readers to the detailed discussion and visualizations provided in the notebook. This evidence supports the log-log transformation as a more effective method for capturing the relationship between carat weight and diamond price across the different quality classes.

### 4.2 COEFFICIENT OF REGRESSIO AND PREDICTION

In the following table, we report the coefficients obtained from the linear regression analyses performed on the log-transformed price and carat data for each diamond quality class. These coefficients, representing the slope and intercept of the regression line for each class, offer insights into the specific relationship between carat weight and price within the distinct quality categories. By analyzing these coefficients, we can understand the varying degrees of impact that carat weight has on the price across the different classes of diamonds.

| | Slope | Intercept |
|---|---|---|
| Ideal | 1.70724 | 8.523274 |
| Premium | 1.657075 | 8.427217 |
| Good | 1.735908 | 8.375691 |
| Very Good | 1.726617 | 8.45256 |
| Fair | 1.494886 | 8.187487 |

Using the coefficients detailed in the table, one can calculate the logarithm of a diamond's price based on the logarithm of its carat weight, according to our regression analysis. This relationship allows for a straightforward estimation of price from carat size within each quality class, utilizing the log-transformed linear model we've established.

While our linear regression model on the log-transformed data presents a compelling fit, it's important to emphasize its applicability primarily for interpolation within the observed data range, rather than for extrapolation to values beyond this range. This caution is underscored by an observation from our analyses: the diamond price data is effectively capped at around $20,000. This upper limit could result from a variety of factors, such as market constraints, limitations inherent to the dataset, among others. Given this observed truncation, and without clear insights into the behavior of diamond prices beyond this threshold, applying our regression model to predict prices for diamonds with significantly larger carat weights would likely yield imprecise estimates. This limitation highlights the importance of understanding the scope and boundaries of our dataset and models when making predictions or drawing conclusions.

## 5. DISCUSSION AND CONCLUSIONS

In conclusion, our exploration of the diamonds dataset has provided valuable insights into the relationship between diamond carat weight and price, alongside the distinctions among different quality classes. Through a combination of distribution fitting, statistical tests, and regression analysis, we've established a nuanced understanding of how these variables interact within the confines of the dataset.

Our findings underscore the relevance of log-normal distribution in fitting the data and the efficacy of log-transformed linear regression for modeling the relationship between carat and price. However, the observed limitation in price data prompts caution against overextending the model's predictive capabilities beyond the current dataset's range.

As we advance, it's crucial to consider these limitations and the specific context of the dataset when applying these models in real-world scenarios. The insights gained lay a solid foundation for further research and application in the diamond market, providing a robust starting point for future investigations into diamond valuation and market dynamics.