A Diamond is Forever

Ashkaan Moinzadeh, Alejandro Reskala Ruiz, Shrey Singhal

Introduction

Diamonds are a unique consumer product. At first glance, there is no apparent utility in owning a diamond. One of the most discernible reasons for owning, or wearing, a diamond lies in its perceived value as a rare and expensive object. Although diamonds have been considered valuable for millennia, their consumption significantly changed in the 20th century when it became customary to give diamond rings as engagement gifts. De Beers, the largest player in the industry, has been using the slogan 'a diamond is forever' to encourage such gifts since 1948.[1] The global sales of diamonds in 2020 reached $68 billion, with the United States being the largest market by contributing $35 billion to total sales.[2] As of today, the United States, China, India and Japan represent the top four markets selling diamonds.

Traditionally, diamonds were sold in jewelry stores. The Diamond District in New York City and the Diamond Quarter in Antwerp have been global centers for diamond trade. With the proliferation of electronic commerce, more diamonds are now being sold on the Internet, allowing for a broader consumer reach. While the perceived value of a diamond may be affected by its price alone, diamonds likewise possess well-defined physical properties. The most commonly quoted of these are what are known as the 4Cs, which represent a diamond's carat weight, cut, color, and clarity.

The goal of our research is to explore the relationship between the physical properties of diamonds and their prices. Applying a set of regression models, we estimate diamond prices as a function its physical attributes. Additionally, our results could have broader implications on the sale of diamonds through online platforms.

Data and Methodology

The data in this study was collected in 2008 from an online diamond retailer[3] and contains the prices and the specifications for more 53,943 diamonds. Each row in the data represents 10 features for a unique diamond - price, carat (weight), cut, color and, clarity, depth, table, x (length), y (width), and z (depth). Of these, 6 are numeric and 4 are non-numeric variables. Moreover, there is no null data in the dataset. We performed all exploration and model building on a 30% sub-sample of the data. The remaining 70% was used to test models and generate the statistics in this report.

Identifying the appropriate sample for our study requires some care. Our exploratory data analysis revealed that over 99% of the diamonds in the dataset were smaller than 2.5 carats. A weight to price distribution plot (Figure 1) confirmed this and also showed us that variance increases with the diamond weight. Moreover, there was a nonlinear relationship in the price and weight of a diamond. We can see that the dispersion or variance of the relationship also increases as carat size increases.

Thus, we limited further analysis to diamonds smaller than 2.5 carats by removing 83 records containing diamonds outside of this range. Moreover, following the recommendations in existing literature[4] to reduce the heteroscedasticity in the dataset, we also transformed the weight and the price of diamonds by taking the natural logarithm of these variables. Figure 2 illustrates the relationship following the transformation.

[1] C. Sullivan. "How diamonds became forever." The New York Times (2013).

[2] De Beers Group of Companies. Diamond Jewellery Demand and Outlook (2016).

[3] https://www.diamondse.info/

[4] W. G. Manning, The logged dependent variable, heteroscedasticity, and the retransformation problem, Journal of Health Economics, vol. 17, no. 3, pp. 283-295, (1998).

Figure 1: Price vs. Diamond Weight (Carats)

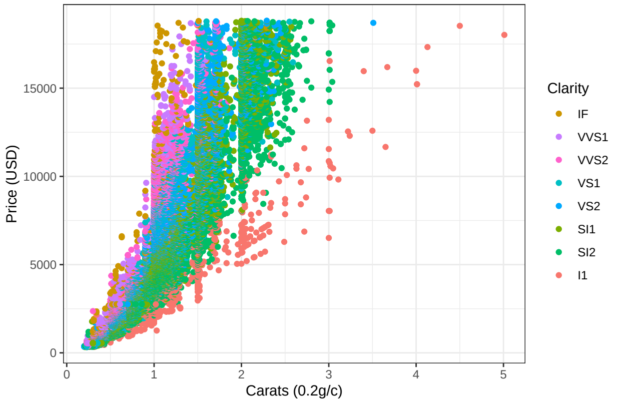

Figure 2. ln(Price) vs. Diamond Weight in ln(Carats)

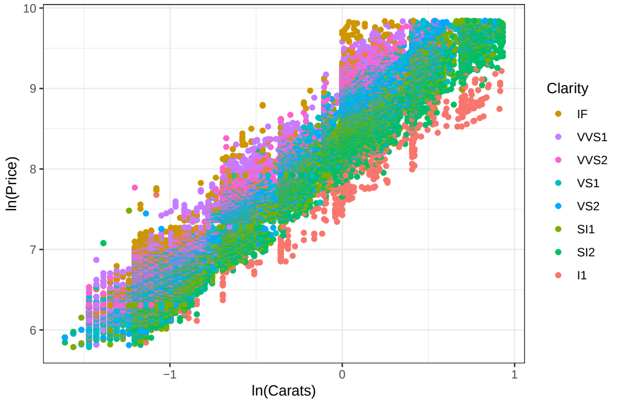

Figure 2 likewise shows that there is a general positive relationship between the weight of a diamond and its price. However, this relationship is non-linear. We believe this could be due to two reasons:

- On the supply side, larger continuous chunks of diamonds without significant flaws are probably harder to find than smaller ones
- On the demand side, customers who buy a less expensive, smaller diamond are probably more sensitive to price than well-to-do buyers. There are fewer customers who can afford diamonds above one carat, and hence we shouldn't expect the market for bigger diamonds to be as competitive as the one for smaller diamonds.

Next, we were interested in analyzing the impact of the remaining 3Cs - clarity, cut, and color on the price of a diamond. For any given carat value, we noted that the diamonds with lower clarity are almost always cheaper than diamonds with better clarity, as demonstrated in Figures 1 and 2.

In contrast, we don't see much variation on cut. Most of the diamonds in the data (40%) are ideal cut, with decreasing numbers of diamonds as cut quality decreases. Finally, color did seem to explain some of the variance in price. While Blue Nile, the leader in the online diamond market, states that the difference between all color grades from D to J are not noticeable to the naked eye, we did observe an association between color differences and prices in the dataset.

Thus, we prepare a model that accounts for all the 4Cs and fit regressions of the form,

$$ln(price) = \beta_0 + \beta_1 \cdot ln(carat) + \beta_2 \cdot clarity + \beta_3 \cdot cut + \beta_4 \cdot color + \mathbf{Z}\gamma$$

where β_1 represents the increase in price of a diamond with increase in its weight, β_2 represents the change in the value with clarity, β_3 reflects the adjustment in a diamond's price with cut, β_4 represents the change in the value with color, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

We also considered the effects of other features such as depth, table, and x-, y- and z-dimensional lengths. However, fitting such models to our exploration set resulted in coefficients that were practically small and non-significant. We omitted these terms to improve the precision and simplicity of our model.

Results

Table 1 shows the results of three representative regressions. The first model is a reduced model with a linear relationship between our output variable (log price) and independent variable (log weight). The second model is representative of the widely known 4Cs for diamond quality.[5] And finally, the third model is a full model that includes the additional features of depth and table. This version also shows us that we could get rid of depth and table, which contribute little overall.

[5] 4Cs of Diamond Quality by GIA — Learn about Diamond Buying — What are the Diamond 4Cs. (2019).

Table 1: Estimated Regressions

	Output Variable: Log (Price of Diamond)		
	(1)	(2)	(3)
Log Weight(Carat)	1.68*** (0.002)	1.88*** (0.002)	1.89*** (0.002)
Clarity IF		1.08*** (0.01)	1.08*** (0.01)
Clarity SI1		0.56*** (0.01)	0.56*** (0.01)
Clarity SI2		0.40*** (0.01)	0.40*** (0.01)
Clarity VS1		0.78*** (0.01)	0.78*** (0.01)
Clarity VS2		0.71*** (0.01)	0.71*** (0.01)
Clarity VVS1		0.99*** (0.01)	0.99*** (0.01)
Clarity VVS2		0.92*** (0.01)	0.92*** (0.01)
Cut Good		0.08*** (0.01)	0.08*** (0.01)
Cut Ideal		0.16*** (0.01)	0.16*** (0.01)
Cut Premium		0.14*** (0.01)	0.14*** (0.01)
Cut Very Good		0.12*** (0.01)	0.12*** (0.01)
Color E		−0.05*** (0.003)	−0.05*** (0.003)
Color F		−0.10*** (0.003)	−0.10*** (0.003)
Color G		−0.16*** (0.003)	−0.16*** (0.003)
Color H		−0.25*** (0.003)	−0.25*** (0.003)
Color I		−0.37*** (0.003)	−0.37*** (0.003)
Color J		−0.51*** (0.004)	−0.51*** (0.004)
Depth			−0.001 (0.001)
Table			−0.0003 (0.0004)
Constant	8.45*** (0.002)	7.89*** (0.01)	7.94*** (0.06)
Weight (Carat)	✓	✓	✓
Clarity/Cut/Color		✓	✓
Additional Features			✓
Observations	37,664	37,664	37,664
R^2	0.93	0.98	0.98
Residual Std. Error	0.26 (df = 37662)	0.13 (df = 37645)	0.13 (df = 37643)
Note:	*p<0.05; **p<0.01; ***p<0.001 HC_1 robust standard errors in parentheses. Additional features are Depth and Table.		

Across all models, the key coefficient for $LogWeight(Carat)$ was highly statistically significant. Point estimates range from 1.68 to 1.89. To provide some sense of scale, by applying the first model, the value of a 1 carat diamond approximates to 5,000 dollars, yet if the weight increased to 2 carats, the price would increase exponential, and the diamond's value will rise to 15,000 dollars.

Furthermore, the analysis confirm that, besides the carat-weight of a diamond, its clarity, cut and color also influence its value. Across models 2 and 3, the coefficients for clarity, cut and color were all highly statistically significant, and likewise improved the precision of predictions by accommodating these features.

Finally, it's important to note the negative sign on the coefficient for various color categories and how the coefficient becomes more negative as we move down the color spectrum (E to J). This matches with the intuition that as a diamond moves away from being colorless towards yellow (E to J), its value should decline.[6]

Limitations

No research is without limitations. Consistent regression estimates require an assumption of independent and identically distributed (iid) observations. Since this exploratory study was conducted on a dataset from a single online retailer, these insights may not generalize to other contexts. For example, it is likely that the consumer experience of shopping for diamonds offline could be very different. Anecdotal evidence suggests that offline diamond business is relatively obscure and involves aspects that are not captured by the compiled datasets.[7] Our study does not yield any information on consumer motivations for purchasing diamonds. For example, consumers could be purchasing diamonds as an investment, and such purchases may be affected by a different set of considerations beyond the diamond attributes that we considered in this study. Further research would be required to test the generalizability of our observations.

Moreover, a dataset that is evenly distributed across all features would create a more generalizable model. Our dataset primarily contains clarity groups SI1, SI2, VS1 and VS2, whereas I1 and IF only make up 1.4% and 3.3% of the data, respectively. However, this population, perhaps, resembles what can be realistically found on the market, and hence, we decided to stick with this dataset and work with this limitation.

Consistent regression estimates also require that the sample group distribution be described by a unique best linear predictor. However, diamonds of similar clarity (or cut and color) show a clustering effect (Figure 2), and each cluster has a unique slope and linear trend over time. Our dataset should be divided across different clusters, and the analysis should be repeated for each cluster. We partially account for this in models 2 and 3, though these models prioritize generalizability across, and not within, clusters.

Conclusion

This study estimates the economic value of a diamond based on its weight and other characteristics. The analysis of the effects of a diamond's physical properties on its price demonstrates that diamond weight generates the greatest effect on price, while color, clarity and shape are, to a lesser extent, also important. Thus, for any luxury retailer, it would be valuable to invest in tools, and subject matter experts, that differentiate across these features. In addition to expanding capabilities to evaluate diamond prices across markets, this model could also reduce consumer search costs for diamond prices, and likewise make it easier for consumers to compare diamond prices given their physical properties, and to do so across different retailers.

In future research, new datasets may be generated to estimate the value of specific types of diamonds. Retailers, and even end-consumers, may wish to know the supply-demand aspect of individual markets, and perhaps even the impact of inflation on prices.

[6] Johnson, M. How Diamond Colour Affects the Value of a Diamond, Serendipity Diamonds (2019).

[7] Y. -K. Ng, Diamonds are a government's best friend: burden-free taxes on goods valued for their values, The American Economic Review, vol. 77, no. 1, pp. 186-191, (1987).