Relationship Between Diamond Characteristics and Its Price

Fall 2022 Section 9 W203 Lab 2 Report - What Makes a Product Successful?
Austin Chen, Jee Park, Saurabh Narain

Introduction

The global diamond jewelry market has seen tremendous growth since the 1880s when the De Beers Group started the diamond mining business. Recent trends show that the global diamond market size is expected to grow at a compound annual rate of 3% from 2020 to 2030, which is valued at USD 89.18 billion in 2019[1]. As the industry will continue to see an increase in demand, we are interested in exploring the relationship between a diamond's characteristics, such as the carat size, dimensions, cut, clarity, and price on the market.

We aim to answer the question: Is there a linear relationship between the carat size, clarity, cut and color of a diamond against its price? The goal is to highlight for diamond retailers the type of relationships that can be observed between diamond characteristics (also known as the 4Cs) and prices, and the potential pricing impact their business decisions may have on consumers. The model will also serve as a way to expose fair market price of a diamond and provide insights on pricing components of a diamond.

The dataset we used is DiamondPrices.csv from Kaggle. It contains approximately 53,900 rows, where each row represents a random, unique diamond. The original data source is from an online jewelry shop which was obtained via web scraping. Given the lack of documentation on the sampling procedure, we recognize there is a risk that the data are not independent and identically distributed. It is, however, documented that each record represents a random diamond, so we proceeded to use this dataset for the analysis.

Conceptualization and Operationalization

To answer our research question, we will establish the conceptualization of a diamond, carat size, clarity, cut, color, perimeter, and price.

A diamond is defined as a standalone stone. While we aim to provide analysis that is to be used by diamond retailer companies, in which case they may be more interested in a fully curated diamond ring that is ready to be sold to consumers, we will scope the model down to a single diamond to minimize potential variability that may come from additional variables.

The 4Cs are: carat, clarity, cut, and color. Carat size represents the unit of weight for stones, where 1 carat is equivalent to 200 milligrams. Carat weight is often considered the most objective grade of the 4Cs[2] and is a critical component of a diamond price because current market behavior shows that it has a positive relationship with price. However, this is just one factor and we aimed to measure the size of the effect compared to other variables. Clarity is the presence or absence of flaws in a diamond. The fewer flaws present on the diamond, the higher the grade. Cut is represented by the proportion and overall finish of the diamond. This does not refer to the shape of the diamond (round, princess, emerald, etc.) but the precision level in which the diamond is cut. Color is defined by the lack of color in a diamond, where a colorless diamond is regarded as the highest quality diamond.

Perimeter is defined as the dimensions of a diamond, in total three components, radiuses of the diameter and height of the diamond.

Price is defined in US dollars for a given diamond. We assume that the price in the dataset was set after the diamond was created, given that each diamond represents a polished stone post-manufacturing.

To operationalize the concepts defined above, we will use the following variables from our dataset: carat, cut, color, clarity, depth, price, x, y, and z.

The concept of a diamond is operationalized as a unique record within the DiamondPrices dataset. All possible values for each variable to be mentioned in this section align with the standard values defined by the American Gem Society (AGS).

[1] https://www.grandviewresearch.com/industry-analysis/diamond-market
[2] https://www.americangemsociety.org/4cs-of-diamonds/

Carat is operationalized by the variable carat, which is well represented to match the conceptualization as the dataset contains records that are within a range of 0.20 - 5.01 carats. Cut is operationalized by the variable cut, which is represented by the following values: Fair, Good, Very Good, Premium, and Ideal. Color is operationalized by the variable color, with the following values: D, E, F, G, H, I, J. The AGS scale spans from D to Z, which means our dataset is restricted to the top tier Colorless (D-F) and Near Colorless (G-J) categories. Clarity is operationalized by the variable clarity, with the following values that indicate flawlessness: I1, IF, SI1, SI2, VS1, VS2, VVS1, VVS2. The AGS scale spans from IF (Flawless) to I3 (Included), and I2 and I3 are not included in our dataset, so we recognize the i.i.d. Data assumption not being met.

Perimeter is operationalized by creating a new variable which is a linear combination of the variables, x, y (radiuses of the diameter), and z (height), by the equation (2x + 2y + z)*4 to normalize into a box shape.

Modeling Decisions

Before creating linear models in R, we explored the dataset by using basic plots to see how variables could potentially relate to each other in contributing to the price of a diamond. We started by splitting the entire dataset into two parts: an exploration set and a confirmation set. The exploration set was used to create all visualizations and initial models. Later in our analysis, we used the confirmation set to evaluate our selected models.

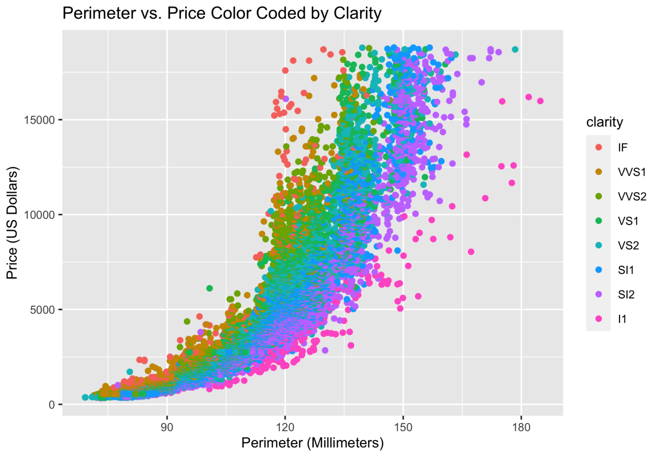
In the initial exploration of the data, we looked at histograms of the x, y, and z dimensions of the diamonds. We noticed that there were several outliers in the dataset in which one or more of the dimensions was 0. This is not possible because diamonds occupy a three-dimensional space, so we started by filtering out all 0s from the entire dataset which resulted in 23 rows being removed.

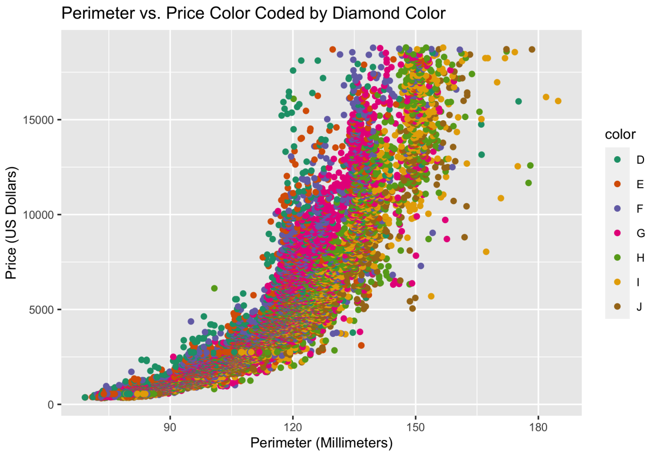
When looking at the distribution of the price, we noticed that it was heavily skewed to the right with diamond prices up to $19,000. This made sense because more diamonds are cheaply priced whereas the highly expensive diamonds are more likely to be rare and infrequent. In an effort to un-skew our outcome variable, we ultimately decided to natural log transform the diamond prices. Thus, all our models predict the log price of diamonds.

When looking at the residual plot of the model, we do see a very slight curvature in the plot. If we include all possible variables that are a part of the dataset, this curvature disappears, however the variables carat and diamond perimeter have very high collinearity.

Since carat and perimeter were collinear with each other, we needed to choose only one. Since perimeter is easier to understand as a first diamond user than carat, we chose perimeter in our final model. We also excluded cut from our model since it did not increase the adjusted R squared significantly and was found to be statistically insignificant at an alpha cutoff of 0.05.

Data Exploration and Visualzation


Perimeter vs. Price Color Coded by Clarity


Perimeter vs. Price Color Coded by Diamond Color

In our exploratory plots, we analyze the relationship between the perimeter of the diamond in millimeters, against the explanatory variable diamond price. The points in the plots are color-coded by the following categorical variables: clarity and color. For perimeter and price, we can see that as the perimeter of the diamond increases and the price increases exponentially based on the shape the points form. This would indicate that we would need to take the log transform of price to account for this relationship.

Additionally, we can see that clarity and color play a significant role in determining the diamond price. In the plot, groups are clustered together based on clarity. For example, I1 is the lowest clarity for a diamond in the dataset and has the cheapest prices based on the plot.

Results

Table 1: Estimated Regressions

	Output Variable: natural log of price in dollars		
	(1)	(2)	(3)
Perimeter (Millimeters)	0.05***	0.05***	0.05***
	(0.0001)	(0.0001)	(0.0001)
Clarity SI2		0.45***	0.43***
		(0.02)	(0.02)
Clarity SI1		0.61***	0.61***
		(0.02)	(0.02)
Clarity VS2		0.74***	0.74***
		(0.02)	(0.02)
Clarity VS1		0.78***	0.81***
		(0.02)	(0.02)
Clarty VVS2		0.91***	0.92***
		(0.02)	(0.02)
Clarity VVS1		0.94***	0.98***
		(0.02)	(0.02)
Clarity IF		1.02***	1.07***
		(0.02)	(0.02)
Color I			0.15***
			(0.01)
Color H			0.29***
			(0.01)
Color G			0.40***
			(0.01)
Color F			0.47***
			(0.01)
Color E			0.50***
			(0.01)
Color D			0.56***
			(0.01)
Constant	2.76***	1.73***	1.09***
	(0.01)	(0.02)	(0.02)
Observations	37,666	37,666	37,666
R^2	0.93	0.95	0.97
Adjusted R^2	0.93	0.95	0.97
Residual Std. Error	0.28 (df = 37664)	0.22 (df = 37657)	0.17 (df = 37651)
Note:			*p<0.1; **p<0.05; ***p<0.01
			HC$_1$ robust standard errors in parentheses

Table 1 shows three different regression models that we considered. Before testing our models on the confirmation set, we explored using the exploration set.

We found the diamond's perimeter to be a statistically significant variable in determining the log price of diamonds. Thus, the relationship between diamond perimeter and log price is positively associated.

We came across diamond clarity and color being statistically significant contributors in determining the log price of diamonds. In general, when the clarity of a diamond is higher, the diamond is likely to be priced higher. Thus, diamond clarity and log price are positively associated. We also observed that the more colorless a diamond is, the higher it is likely to be priced. Diamonds that are nearly colorless, were generally priced cheaper in comparison to diamonds that are colorless. Thus, diamond color and price are positively associated.

After exponentiating the coefficients from our model, we can make a few meaningful interpretations. For each increase in millimeter of diamond perimeter, the diamond price in dollars increased, on average, by about 5%. A diamond with clarity IF (flawless) is likely to be about 190% more expensive in dollars on average compared to a diamond with clarity I3 (included). A diamond with color D (colorless) is likely to be about 74% more expensive in dollars on average compared to a diamond with color J (nearly colorless).

Limitations

Since the dataset contains approximately 54,000 entries, we can apply the large-scale linear model assumptions. All assumptions were met with one slight exception. In the residual plot for the final model, we noticed a slight bend that could not be accounted for when reasonably including other variables in our model. This means that there could be other contributing factors that we are not factoring in that contributed to this slight bend in the residual plot.

Additionally, the dataset excludes I2 and I3 levels for clarity, so it does not account for the full range of possible diamonds and their prices. This raises a potential risk in the model where the coefficient for clarity may be over or understated than the true value. We recognize that the analysis, therefore, only extends to a portion of the lower-end, "Included" group of diamonds and would look for a more comprehensive dataset in the future.

Another limitation is the lack of information on the manufacturing origin of the diamond and the country in which the diamond was priced. For example, if the dataset was pulled from the United States only, the listed prices may not be a realistic global price, assuming Europe, Asia and other parts of the world may price diamonds differently.

We are also not taking the diamond production methodology into account, which could influence the results. Lab-grown diamonds are significantly cheaper than natural diamonds given the production process is low-cost for the former. We cannot confirm whether the dataset includes both kinds of diamonds, so the interpretation of the model results holds under the assumption that this represents any diamond.

Another limitation is that we do not know the times that the prices of each of the diamonds were collected. If the diamond prices were scraped at roughly the same time, then we would not need to factor in inflation into our model. Otherwise, we believe having the time that the individual entries were collected would serve useful in our model.

Conclusion

The statistical report estimated the diamond price based on the perimeter, clarity, and color of the diamond. Initial data exploration suggested a positive relationship between the predictor variables and the price of the diamond, and the statistical model confirmed our hypothesis.

If we were to improve our statistical report, we would like to gather the origins of the diamonds to provide a tool for consumers based on their geographical region to provide an accurate price of a diamond for them. This will help first-time diamond buyers worldwide purchase a diamond at a fair price based on their region. We would also explore non-linear relationships between the variables in the current dataset because we do believe that more complicated interaction terms can be made with the existing data.

All code for the project can be found in this GitHub repo: https://github.com/mids-w203/lab2-austin-jee-saurabh