Estimating the Impact of Synthetic Diamond's Weight on Sale Price

Datasci 203 Lab 2 - Research report for Acme Synthetic Diamond Company

Ivy Chan, Jonathan Hodges, Dipika Kumar, Christian Lee

December 6, 2022

Contents

1 Introduction **1**

2 Data and Methodology **1**

3 Results **2**

4 Limitations **3**

5 Conclusion **4**

1 Introduction

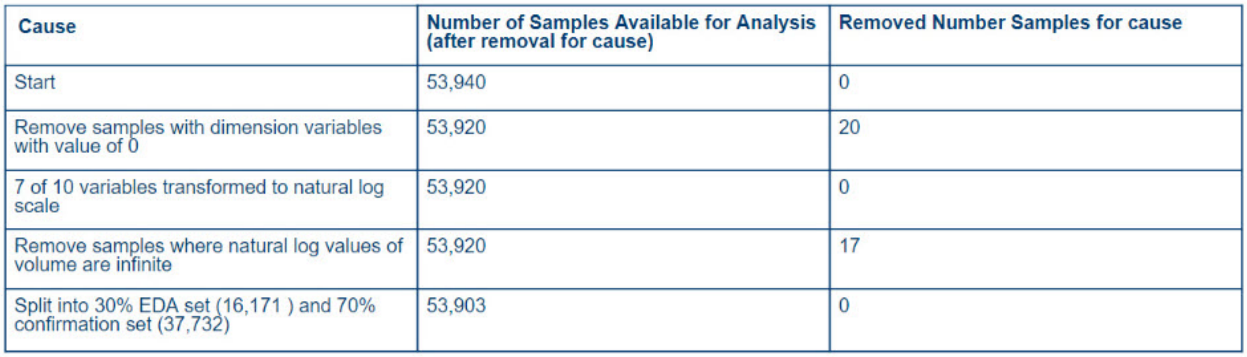
New technological advances in synthetic diamond manufacturing have led to increased carats of the produced synthetic diamonds. Acme Synthetic Diamond Company is still using the older manufacturing process. Since upgrading the new equipment will be expensive, they have to decide if it is economically viable. Generally, diamonds with larger weights or carat values sell for higher prices, a more precise data analysis is required to justify such a large investment.

This study estimates how large of a factor the carats of synthetic diamonds are on the sale price of the diamond. We leverage synthetic diamond sales observations with the sold diamonds' characteristics, including carat, color, clarity, cut, length, width, and depth. Applying a set of regression models, we estimate how much the synthetic diamond weight in carats influences price with respect to other factors.

2 Data and Methodology

The data in this study comes from the diamonds dataset on Kaggle [1] It was made publicly available by Abhijit Singh in 2021.The data includes 53,940 observations of synthetic diamond sales with 10 variables. We transformed 7 of the 10 variables to the log scale to remove skewness and make them more symmetrical with more normal distributions. We have determined that records with a value of zero in columns volume, length, width, depth, and table were to be removed since it is impossible for a physical dimension of a diamond to be an absolute zero. This filtering removes 20 observations and leaves us with 53,920 observations. In addition, we removed 17 observations that have the infinite natural log volume value and that leaves us with 53,903 observationsin our data set.

We assigned 30% of the data to the exploration set, 16,171 observations, and the remaining 70% of the data to the confirmation set, which is 37,732 observations. The large observation sizes are sufficient for the central limit theorem (CLT) to hold. The exploration set was used to inspect the data's trend and build models, while the confirmation set was used to test our model on new data.



Cause	Number of Samples Available for Analysis (after removal for cause)	Removed Number Samples for cause
Start	53,940	0
Remove samples with dimension variables with value of 0	53,920	20
7 of 10 variables transformed to natural log scale	53,920	0
Remove samples where natural log values of volume are infinite	53,920	17
Split into 30% EDA set (16,171) and 70% confirmation set (37,732)	53,903	0

Table 1: Accounting Table

To predict the price model with the different diamond features, we had to operationalize varying features. The price is metric and was operationalized as the response variable of the model, while the carat, volume, clarity, and color were used as predictor variables. The volume variable, a metric feature, is made of three different features in the data set, the multiplication of the width_in_mm, height_in_mm, and depth_in_mm of the diamond. The rationale for multiplying the three metric features to be one feature denoted as volume is to simplify the interpretation of the model. The clarity, color, and cut are categorical variables that were operationalized by being hot encoded to ordinal features. Having these variables as ordinals defines the categorical variables to their price worth based on their physical properties and will generate a more precise model.

[1] EDA & Applying Multiple linear regression. "https://www.kaggle.com/code/abhijit10singh/eda-applying-multiple-linear-regression/data/" (2021).

Utilizing the exploration set and going through the EDA process, plotting the different variables against price, generating different histograms, and running some models with the exploration set, we determined a natural log of of skewed variables depth_in_percent, depth_in_mm, carat, price, width_in_mm, length_in_mm, table_in_percent would produce a better model.

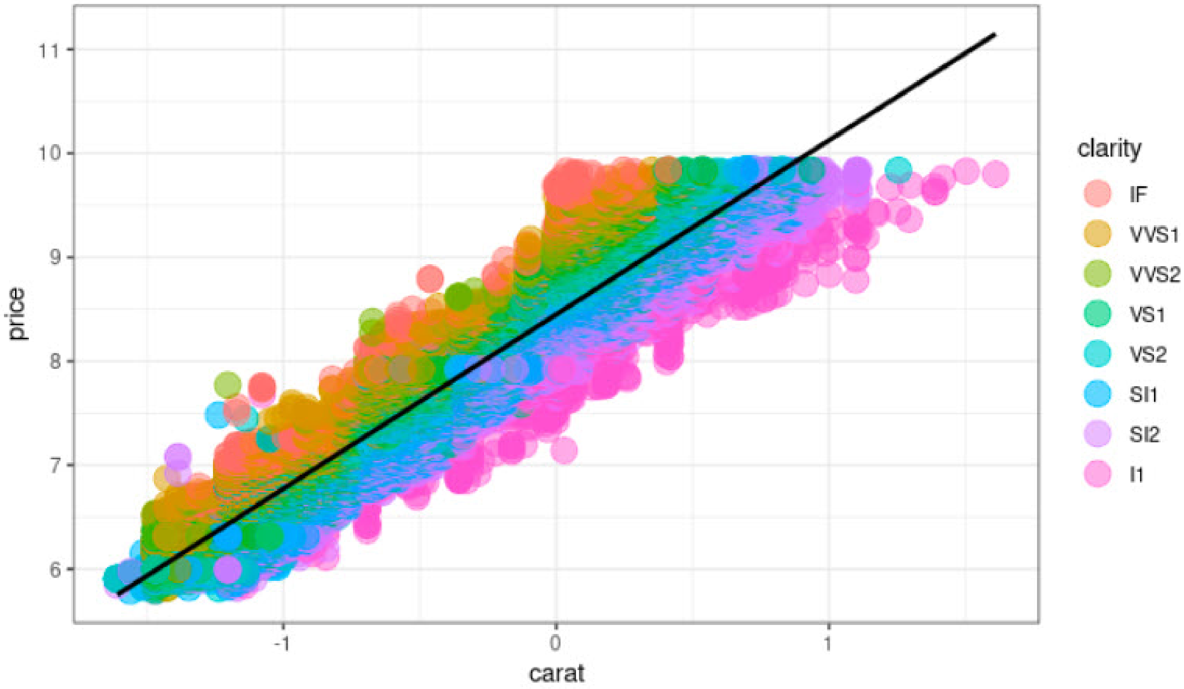


Figure 1: Synthetic Diamond Price Based on Carat and Clarity

$$\widehat{price} = \beta_0 + \beta_1 C + \mathbf{Z}\gamma$$

We are interested in estimating the sale price of synthetic diamonds based on the weight of the diamonds in carats. β_0 is the intercept, and β_1 is the expected change in price based on per unit change in C (carats). $\mathbf{Z}$ is a row vector of additional covariates and γ is a column vector of coefficients.

3 Results

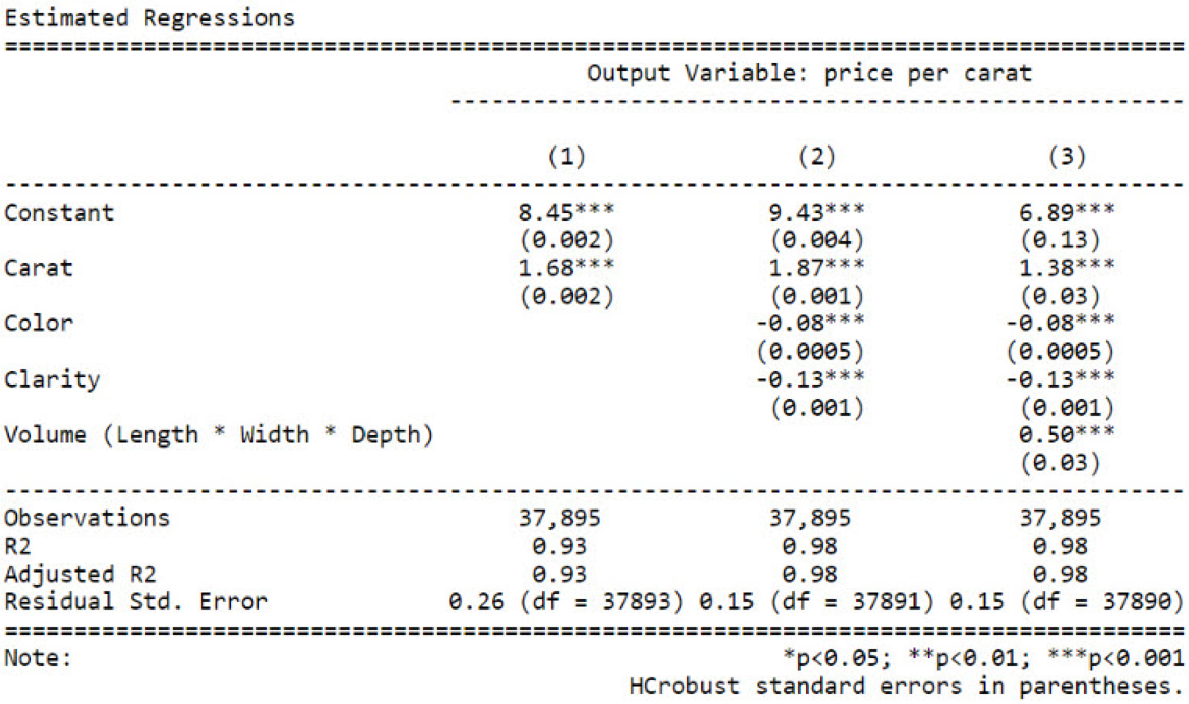
The stargazer table shows the results of the three log-log regressions, and model 1 is our baseline model. Across all linear models, the key coefficient was carats and was highly and statistically significant. It is a positive coefficient with point estimates ranging from 1.38 to 1.87. Model 2 has the highest positive coefficient for carats of 1.87, while models 1 and 3 also have the positive coefficient for carats of 1.68 and 1.38 respectively. We notice in model 3, volume is highly correlated with carat and takes away some explainability of carat, so we decided model 2 is best suited for our estimations. Applying model 2 with the point estimates of 1.87, we interpreted the result that a 10% increase in carats results in a 19.51% increase in price, and a 20% increase in carats is a 40.63% increase in price [2]. As a result, the new equipment with the new process that increases the carats has both practical and statistical significance to the synthetic diamond sales price.

In model 2, we include the independent variables of color and clarity in the linear models. We want to find out if these two variables have any effect on the industrial diamond price. The regression table results

[2] Log Transformation: Purpose and Interpretation. "https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9"

show that both variables are highly and statistically significant and have a negative coefficient. Therefore, it confirms that they do affect the industrial diamond price.

The main difference between models 2 and 3 is the additional independent variable of volume. In model 3, we added an interaction term of volume, which is the product of industrial diamond length, width, and depth.The regression table shows that it is also statistically significant.



```
Estimated Regressions
================================================================================
                                          Output Variable: price per carat
                                ------------------------------------------------

                                     (1)              (2)              (3)
--------------------------------------------------------------------------------
Constant                           8.45***          9.43***          6.89***
                                  (0.002)          (0.004)          (0.13)
Carat                              1.68***          1.87***          1.38***
                                  (0.002)          (0.001)          (0.03)
Color                                              -0.08***         -0.08***
                                                  (0.0005)         (0.0005)
Clarity                                            -0.13***         -0.13***
                                                   (0.001)          (0.001)
Volume (Length * Width * Depth)                                      0.50***
                                                                    (0.03)
--------------------------------------------------------------------------------
Observations                       37,895           37,895           37,895
R2                                  0.93             0.98             0.98
Adjusted R2                         0.93             0.98             0.98
Residual Std. Error         0.26 (df = 37893) 0.15 (df = 37891) 0.15 (df = 37890)
================================================================================
Note:                                             *p<0.05; **p<0.01; ***p<0.001
                                               HCrobust standard errors in parentheses.
```

Table 2: Stargazer Table

4 Limitations

The model has over fifty-seven thousand data points which satisfy the large sample assumption model and allows us to apply a less stringent OLS regression with fewer assumptions. The first assumption is independence and identically distributed (IID) Data. The data could be assumed independent since the data was collected randomly; however, the geographic sampling location of the data is unknown. This could lead to the clustering of information on diamond prices, and the geographic location of buying a diamond could vary the price. The data is identically distributed since we did not remove a significant amount of data (less than 0.04% of data) before or after creating the model and the data points come from the same probability distribution. Please refer to the brief description section for data removal before modeling.

The other aspect to investigate is the variance of the tails. We plotted the histograms of the metric variables (model, price, carat, and volume) and noticed non-normal distributions. The distributions seen were bimodal and skewed, which could lead to a bias in the model's estimate. There seems to be some conflict with the large sample assumption of a unique best linear predictor (BLP), given the distribution of price on carat does display heavy tails. This does not mean the predictor is inaccurate in predicting price; however, we can not assume there is a unique BLP.

Omitted variables do not interact with the key variable in the true model in the classic omitted variables framework. There may be unknown variables that may bias our estimates that could bias our estimate positively or negatively. For example, the labor costs leading to the sale of diamonds could be higher in certain economic conditions, driving up the cost of goods sold and, ultimately, the diamond price. There are such variables that we cannot account for, but it could contribute to our omitted variable bias.

5 Conclusion

In our models, carat is the most important feature in determining the price of a synthetic diamond with the largest coefficient in the regression tables. However, the regression tables also showed there are other predictors of note in volume, clarity and color, potentially adding bias to our estimates that showed 10% and 20% increases in carats leads to 19.51% and 40.63% increases of sale price respectively. That said, with 0.98 for R-squared we have enough confidence to recommend that Acme Synthetic Diamond Company should upgrade to the new equipment. We estimate the larger diamonds produced in carats will bring in more sales revenue, allowing them to pay off the new equipment investment in about three years.

In future research, we may want to collect additional data to estimate the diamond price further. Possible new variables include seasonality, Gemological Institute of America (GIA) certification, the energy used in the manufacturing process, the diamond shape, the jewelry type, or other factors. We hope our research can help Acme Synthetic Diamond Company produce beautiful and profitable industrial diamonds for generations of customers to come.