Multivariate Linear Regression Analysis on the Sales Price of Cubic Zirconia

---

Executive Summary

Study Problem & Hypothesis

The analysis aims to determine the attributes of cubic zirconia that significantly influence the sales price of the gemstone. Cubic zirconia, which is a lab-grown diamond simulant made of zirconium dioxide instead of pure carbon, has close similarities to that of a diamond in that both gemstones are colorless and give off a similar shine under light (Antolik, 2022). Given the similarity between these gemstones, evaluating these attributes/factors can further be expanded to explore those of cubic zirconia. Furthermore, understanding how these attributes may influence the overall value of the gemstone also gives better insights into the potential profitability of sales.

Null Hypothesis: There is no significant association between the cubic zirconia attributes and the sales price.

Alternative Hypothesis: There is a significant association between at least one of the cubic zirconia attributes and the sales price.

Data Analysis Process

The data was obtained from Kaggle and is authored by a collective group called the Co-learning Lounge. The dataset contained 26,967 records and 11 variables. The variables comprised the attributes of cubic zirconia such as carat weight, cut quality, color grade, clarity, gemstone depth, table size, length, width, height in mm, and price.

The data was cleaned, resolving any issues of missing values, duplicates, and outliers. Any record that missing values, duplicates, and extreme outliers deemed physically impossible for the gemstone's size were removed.

Once the data was cleaned, it was prepared before any modeling was performed. The categorical variables were all ordinal, so each was numerically label-encoded in alignment with the data dictionary. Once data preparations were complete, the data was split into training and testing sets using an 80/20 split.

The sales price was regressed against all of the other variables of the dataset. During the regression analysis, the assumptions of multiple linear regression were checked. These assumptions included checking for multicollinearity, homoscedasticity of the residuals, and multivariate normality (Bobbitt, 2021). Linear regression models were trained on the data and checked against the assumptions to validate the statistical significance of the features and the models themselves.

Lastly, a model evaluation metric, using root mean squared error (RMSE), was used to determine the model's predictive accuracy on the testing data.

Study Findings

Once the data was modeled, it was found that there were two significant factors: carat weight and color grade. The optimized model involved modifying both the target and features using natural logarithmic transformations. As seen below, the features and the model were statistically significant as the p-values were less than .05, of which the significance level was set at.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.946
Model:	OLS	**Adj. R-squared:**	0.946
Method:	Least Squares	**F-statistic:**	1.825e+05
Date:	Thu, 07 Nov 2024	**Prob (F-statistic):**	0.00
Time:	11:53:04	**Log-Likelihood:**	355.26
No. Observations:	20981	**AIC:**	-704.5
Df Residuals:	20978	**BIC:**	-680.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.1464	0.005	1666.578	0.000	8.137	8.156
carat	1.7320	0.003	595.614	0.000	1.726	1.738
color	0.2352	0.003	67.790	0.000	0.228	0.242

Omnibus:	433.091	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	903.182
Skew:	-0.085	**Prob(JB):**	7.52e-197
Kurtosis:	4.002	**Cond. No.**	6.48

The equation of the final model is seen below. Keeping all other variables constant, increasing only carat weight or color grade, individually, by 1% would increase the sales price by 1.73% and 0.24%, respectively.

$$\ln(price) = 8.15 + 1.73 \cdot \ln(carat\ weight) + 0.24 \cdot \ln(color\ grade)$$

Lastly, an error metric was examined to determine how well the model performs and can generalize to out-of-sample data. Per calculations, the RMSE for the training data was 0.2379, and for the testing data, 0.2383. Due to the error values closely mirroring one another, the model can perform well on out-of-sample data.

Study Limitations

A limitation of the research involves the analytical technique used, linear regression. Linear regression makes strong assumptions about the linear relationship between the variables, which may not reflect how the variables are perceived. This assumption strongly influences how the model is trained and the predictions that come with it. If underlying relationships are not truly linear, the regression model is not representative and may lead to poor estimates.

Recommended Actions

Based on the results of this analysis, it is recommended that the model can be used for predictive measures. Further testing could be done further to validate its predictive accuracy on other data before deployment. For future dataset studies, instead of focusing on impactful predictors, other studies could instead focus on generating the best predictive model using different techniques such as random forests or regularization techniques like lasso or ridge regression. Another study point would be employing clustering techniques to better understand the characteristics that make up the different groupings.

Expected Benefits

A benefit of this study was the ability to reduce the complexity of the dataset, from 9 features down to 2, to determine only those attributes that significantly influence sales price.

Furthermore, another beneficial outcome was modeling a productive linear regression model. The simplicity of this model allows it to better generalize on new data. The model can be used to effectively predict cubic zirconia sales prices, allowing entities to determine the profitability of the gemstone.

Sources

Antolik, C. B. (2022, October 18). *Cubic Zirconia Vs. Diamonds*. International Gem Society. https://www.gemsociety.org/article/cubic-zirconia-vs-diamonds/

Bobbitt, Z. (2021, November 16). *The Five Assumptions of Multiple Linear Regression*. Statology. https://www.statology.org/multiple-linear-regression-assumptions/