A Multivariate Analysis Approach to Diamonds' Pricing Using Dummy Variables in SPSS

Ana Rita Antunes[1(✉)], Cláudia Buga Buga[2], Daniel Coutinho Costa[3], José Grilo[3], Ana Cristina Braga[1], and Lino A. Costa[1]

[1] ALGORITMI Center, University of Minho, 4710-057 Braga, Portugal
id9069@alunos.uminho.pt, {acb,lac}@dps.uminho.pt
[2] DTx Digital Transformation CoLab, University of Minho, 4800-058 Guimarães, Portugal
id9400@alunos.uminho.pt
[3] Center for Micro-Electro Mechanical Systems (CMEMS), University of Minho, 4800-058 Guimarães, Portugal
{id9089,id8983}@alunos.uminho.pt

Abstract. In the present article, a distinctive methodology to predict the price of diamonds is proposed. To do so, a vast dataset was used encompassing empirical data relative to variables commonly used to assess the commercial value of diamonds. The selected dataset was retrieved from the Kaggle website [1] and includes diamonds' physical properties along with their respective price. Therefore, a data analysis based on Analysis of Variance (ANOVA) was conducted to study the diamonds' characteristics that determine their prices. It was found that the weight of the diamond (*carat*) has an impact on diamonds' price. For this reason, diamonds' price per carat was considered as a new dependent variable. Moreover, the variables diamonds' *clarity* and the width of the top of the diamond (*table*) affect the dependent variable. After applying the stepwise regression methods, it was found that the variables related to the width of the largest section of the diamond (Y), and *table* were the least significant ones. Moreover, both backward and forward selection led to the same result in terms of the predictive model. All the residuals' assumptions were validated. The adjusted coefficient of determination was 88.3%. Since multicollinearity effects can exist between the independent variables, Principal Components Analysis (PCA) can be used, as future work, to eliminate these effects.

Keywords: Diamonds' price · ANOVA · Linear regression · Stepwise methods · Dummy variables · SPSS

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12952, pp. 609–623, 2021.
https://doi.org/10.1007/978-3-030-86973-1_43

1 Introduction

It is common knowledge that diamonds are valuable items and often associated with elevated market values and luxury items. However, their applications go far beyond wealth and ostentation. Given their extreme mechanical hardness, stability, optical transparency, and outstanding thermal conductivity they can be applied as a valuable feedstock material in several fields [25]. Despite being a natural occurring carbon-based mineral, that can be isolated from crude oil, natural gas, and hydrocarbon materials, such as Adamantane and Diamantane, diamonds have been a research topic for a wide range of areas including geology [13], biology [25], economy [31], material and manufacturing [16]. Even though the price is influenced by three market segments (industrial, jewellery, and investment), in the past it was stated that diamond's price fluctuation is highly correlated with disposable income and inflation rate [4,24]. In 1995, the introduction of synthetic diamonds [4] offered a more economically accessible manufacturing option to the production of diamonds, which were similar in properties and aesthetics, but customizable and limited in terms of dimensions. Nonetheless, this created a steep increase in the diamond price, as well as an enlarged demand for research and development [24] as the list of diamond consumer industries grew larger [5,26]. Recently, R. Bedoui et al. conducted studies on the diamond investing potential [3], concluding that even though diamond tends not to function as a safe haven and hedge for USD (United States Dollar) movement, it should still be used as an effective diversified to mitigate market volatility.

As previously mentioned, the main goal is to find the factors that influence the diamonds' price and the relationship between the independent variables with diamonds' price.

This paper is organized as follows. In Sect. 2, the methodology used in this work is described. The case study and the results obtained are presented and discussed in Sect. 3. Finally, some conclusions and future work are drawn in Sect. 4.

2 Methodology

In this section, the statistical techniques used to analyse the dataset are described, namely analysis of variance, multiple linear regression and variables selection methods.

2.1 Analysis of Variance

The Analysis of Variance (ANOVA) technique uses the variance to compare means and it is an inferential test that can be used when there are factors to be explored [22,23]. Therefore, ANOVA analyzes, through the sample variance, if there are differences between the mean of different levels in each factor or treatment.

In order to accurately assess how the levels differ, for example, to check if in each level there are in fact significant differences, it is necessary to use post hoc analysis, like Tukey's procedure [23,28]. Furthermore, different research areas use ANOVA to answer the research questions to extract more information about the data in the study [22].

Thus, One-way ANOVA is a technique that compares the means of more than two groups (independent variables). It takes into account the variability between groups and the variability within groups, where an F-statistic is produced. It is required that the dependent variable follows a normal distribution, the samples must be independent, the errors must be also independent and normally distributed. Furthermore, the variance must be equal between the groups, also known as homoscedasticity. If the normality assumption is violated, ANOVA is still robust [18,27].

Using ANOVA, it is expected to understand if there are differences between the diamonds' price mean of the different levels of each qualitative variable. This analysis can provide useful information about which qualitative variables can be a potential factor to influence the diamonds' price.

2.2 Multiple Linear Regression

Regression analysis is a statistical method that allows building a model for one quantitative dependent variable (y_i) as a function of one or more other independent variables ($x_i = 1, ..., p$). Its purpose is to infer the causal relationship of data obtained from correlational or non-experimental research. Multiple linear regression is employed to assess the significance of the independent variables on the diamond's price. Thus, the linear model is defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where β_0 represents the additive constant; $\beta_1, \beta_2, \ldots, \beta_p$ the coefficients of the independent variables; ϵ_i the random error component of the model.

Model validation is performed by analyzing the Gauss Markov conditions for residuals [8], which includes verifying zero mean condition and homoscedasticity of the errors as well as normality and independence hypothesis. Additionally, influence diagnostics is carried out through outlier detection, Leverage values, and Cook's distance. Leverage values (h_i) are calculated from the values of the independent variable, which assesses the distance of observed response from the point average of all independent variables. In this context, a Leverage value is identified as an outlier under the following condition: $h_i > 2(p+1)$ [17]. Cook's distance (D_i) is the squared distance between the least square estimated based on all observations n points $\hat{\beta}$ and the estimate obtained from eliminating the ith point $\hat{\beta}_{(i)}$ and it is used to measure the influence of removing an observation [20]. Such distance measure can be expressed as,

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\beta}_{(i)} - \hat{\beta}_{(i)})}{pMSE}$$

Using regression analysis, it is expected to establish a relationship between the diamonds' price and its independent variables through means of a linear function. From the resulting analysis, conclusions can be drawn regarding the significance of the independent variables towards the dependent variables.

2.3 Variables Selection

To select the variables that should enter the predictive models, stepwise selection methods were frequently used, thanks to their reliability [14]. Simultaneously, the value of Akaike Information Criterion (AIC), associated with the obtained models, can be used to estimate the prediction error and analyze the quality of several predictive models comparatively [12,15].

The stepwise methods are used to limit the number of possible regressions evaluated by the software during the process of creating a multivariate linear regression. They evaluate a limited array of independent variables at a time and define which ones fit the dependent variable data better. Depending on the way those independent variables are selected they can be categorized into the forward selection, backward elimination, or stepwise regression method [19].

To select the independent variables for the final model through the stepwise regression method, the selection process is repeated until a maximum number of steps, or until none of the remaining variables is eligible to be entered or removed [14]. The stepwise method is based upon two defining algorithms that rely on probability (P) and F values, and all eligible variables are considered for removal and entry at each step [14].

In the case of the forward selection, only the entry phase of the stepwise method is performed. The test begins with no predictor variables and they are progressively added one at a time. The first added variable is the independent variable that presents the highest correlation with the dependent variable. Then, the variable with the highest correlation with the residuals from the previous model is chosen and this procedure is repeated until the required F ratio is met [11]. During this process, the coefficient of determination (R^2) should be monitored as a way of uncovering how much predictive capability each step iteration provides [21]. On the other hand, the backward stepwise regression begins with the entire set of independent variables and progressively eliminates them using the removal algorithm of the stepwise regression, until a satisfactory model is achieved [19].

Despite the importance of these methods for the creation and simplification of predictive models, some limitations are associated, which often lead to an unsatisfactory final model. This happens because the act of selecting a subset of independent variables, relying solely on mathematical algorithms, might cause the removal of variables that have critical importance to the study. One example occurs when two variables present almost the same predictive ability. Faced with this issue, the software might end up adding or removing the wrong variable, since it lacks the previous critical knowledge held by the user [21]. This means that even though these methods are very convenient, especially in predictive studies that encompass a large array of independent variables, ignoring the

authors' knowledge about the variables' importance might lead to poor and not generalizable models [29]. Hence, the AIC value can serve as an important tool to further evaluate the resulting models from a comparative stance. This value helps to predict which fitted candidate model best suits the data [7].

In this work, stepwise procedures are used to develop an encompassing multivariate predictive model. Nonetheless, the added/removed variables will be critically analyzed at each step and their predictive value will be assessed taking into account all the information gathered, specifically during the descriptive analysis, ANOVA, and correlation steps, that will map the importance of each independent variable and uncover relationships between them. Moreover, after obtaining the models through the different stepwise regression methods, the model with the smallest associated AIC value will be selected.

3 Case Study

The presented methodology was applied to the diamonds' dataset. Firstly, a descriptive analysis is performed. Then, the relationship between the independent qualitative variables, with the diamonds' price, is tested using ANOVA. For the independent quantitative variables it was used Spearman's correlation. Multiple linear regression encompasses the model construction routine and the analysis of a multiple predictive regression model for the prediction for the diamonds' price based on its attributes. All results were obtained using IBM® SPSS® Statistics, version 27.

3.1 Descriptive Analysis

The diamonds' dataset [1] includes several variables concerning the diamonds characteristics and price. In Table 1, all variables and the respective description are presented.

Table 1. Diamonds dataset information [6,10].

	Name	Description	Measure
Independent	Carat	Weight of the diamond	Quantitative
	Clarity	A measurement of how clear the diamond is	Qualitative
	Color	Diamond color (colorless and near colorless)	Qualitative
	Cut	Quality of the cut	Qualitative
	Table	Width of top of the diamond relative to the widest point	Quantitative
	X	Length (in mm)	Quantitative
	Y	Width (in mm)	Quantitative
	Z	Depth (in mm)	Quantitative
Dependent	Price	Diamonds' price (in US dollars)	Quantitative

In this dataset, there are nearly 54,000 rows of data sampling regarding the diamonds' information. Since the number of observations is too high, 2% of them were randomly selected (approximately 1,073 diamonds). This was done due to the high computational cost involved when handling this much data. The selected sample size guarantees a good representation of the sample for the analysis and predictive model construction.

Carat, *cut*, *color*, and *clarity* are usually known as the four Cs and are the most commonly laboratory-studied variables when it comes to diamonds' price estimation [9,30]. The *clarity* variable has five groups: seriously flawed, slightly included, very slightly included, very slightly included, and internally flawless. The color variable has near colorless and colorless groups and the cut variable has fair, good, very good, premium, and ideal levels. *Cut*, *color*, and *clarity* groups are not balanced, since it was extracted only 2% of the data sampling.

After a brief data analysis, twenty outliers were found and removed. Additionally, the variable *table* was coded into two different levels since it only has 25 different values, where level one represents the *table* values less than or equal to 57 (*table* mean). Otherwise, the measurements belong to the second level.

To assure the symmetry of the distribution of the quantitative variables, another transformation was required. Thus, the natural logarithm was applied to these variables, since their distribution was asymmetrical. Moreover, it is required that the residuals follow a normal distribution. Using this transformation can help to validate this assumption. Regardless of this transformation, these variables will be referred to the "new" variables *log(X)* as *LX*, *log(Y)* as *LY*, and *log(Z)* as *LZ*. It was found that the *carat* variable has a considerable impact on the diamonds' price. Hence, the dependent variable was changed to diamonds' price per carat (*dpc*). To assure symmetric distribution of the *dpc* variable, the logarithmic transformation was also applied (*ldpc*).

3.2 Analysis of Variance

In order to evaluate which qualitative variables affect the dependent variable (*ldpc*), one-way ANOVA was used considering four factors: *clarity*, *color*, *cut*, and *table*.

The hypotheses being tested are similar where it is intended to evaluate if the factors affect the diamonds' price per carat, individually. Considering $i = \{cut, color, clarity, table\}$, the hypotheses are:

H_0: The $factor_i$ does not significantly affect the diamonds' price per carat;
H_1: The $factor_i$ significantly affects the diamonds' price per carat.

In Table 2 the results using one-way ANOVA are presented since the dependent variable follows a normal distribution.

The variables *table* and *clarity* affect the diamonds' price per carat since the $p - value$ is less than the level of significance ($\alpha = 0.05$). In contrast, the variables *cut* and *color* do not affect the diamonds' price per carat, where the $p - value$ is greater than the level of significance.

Table 2. One-way ANOVA results.

Source	Df	F	P-value
Table	1	8.901	0.003
Cut	4	0.558	0.693
Color	1	1.650	0.199
Clarity	4	2.988	0.018

In order to evaluate the homogeneity of variance and randomness of residuals, Levene's test was performed. The $p - value$ is equal to 0.061, which means the variance of the diamonds' price per carat is equal across groups. Therefore, the assumption for homoscedasticity is validated. Figure 1 presents the Normal Q-Q Plot to evaluate the normality assumption and most of the values are according to the diagonal. That means, the residuals follow a normal distribution. The same conclusion can be made using the Kolmogorov-Smirnov test of normality, in which the $p - value$ is equal to 0.067. In addition, the residuals' mean is equal to zero. All assumptions for the residuals are satisfied.

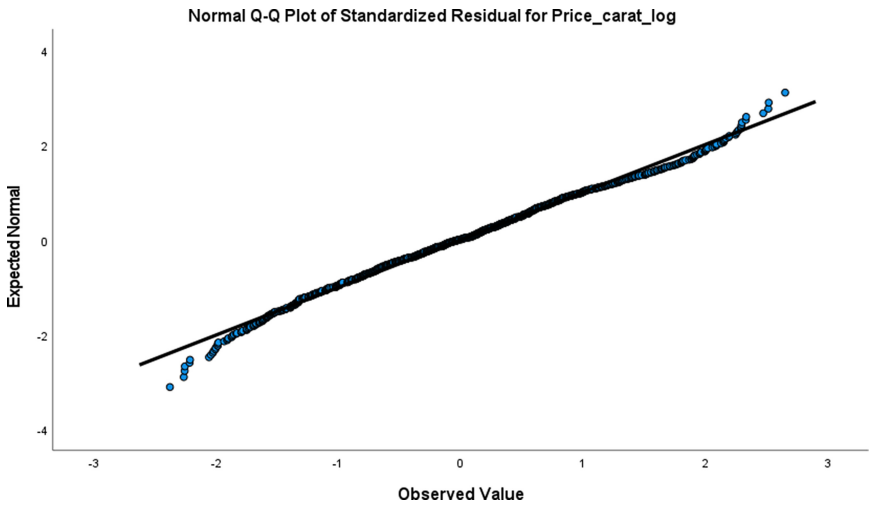

Fig. 1. Normal Q-Q Plot for the ANOVA residuals.

The Tukey HSD method was used for multiple comparisons between groups. It was only necessary to produce multiple comparisons for the variable *clarity* since it was more than two groups. Therefore, the groups seriously flawed and sightly inclusions have differences in the diamonds' price per carat ($p - value = 0.009$). The groups seriously flawed and very, very slightly included also have differences in the dependent variable ($p - value = 0.028$).

Figure 2 and Fig. 3 allow visualizing the average values for the levels of factors *clarity* and *table*, respectively. It is possible to conclude that the group seriously flawed has the lowest impact on the diamonds' price per carat when compared with the other *clarity* groups. Conversely, the second *table* group ($Table > 57$) has a greater impact, on the diamonds' price per carat, than the first level ($Table <= 57$).

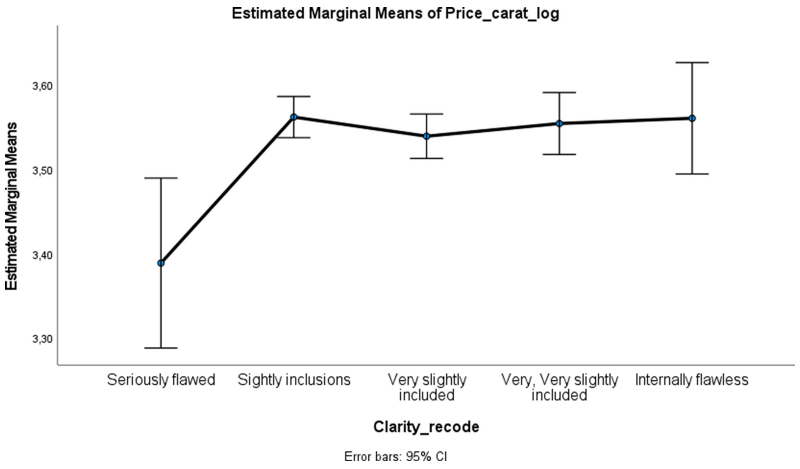

Fig. 2. Profile plot for clarity levels.

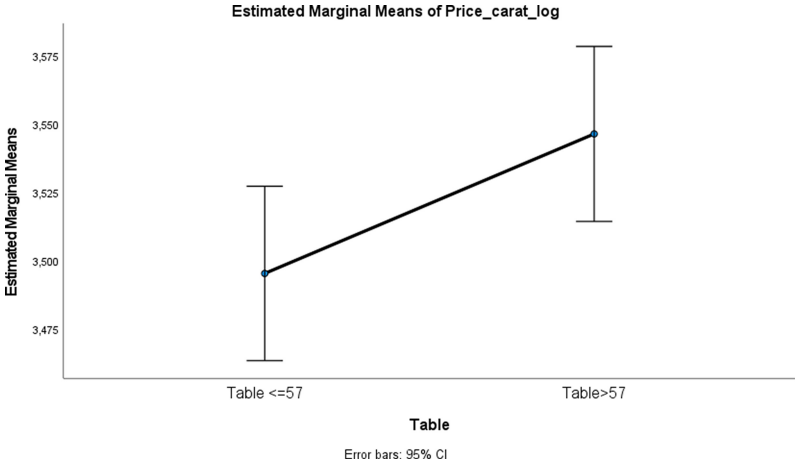

Fig. 3. Profile plot for table levels.

To analyze the relationship between the independent quantitative variables (measures *LX*, *LY*, and *LZ*) and the dependent variable of the sample, the diamonds' price, correlations were computed. Firstly, the normality of the variables

ldpc and *LX*, *LY*, and *LZ* was assessed, by the Kolmogorov-Smirnov test. All obtained $p - values$ were less than 0.001 which allows the rejection of the null hypothesis. Thus, the variables do not follow a normal distribution. Hence, the non-parametric Spearman's correlation coefficients (r_s) were computed for the diamonds' dimensions and the diamonds' price (Table 3).

Table 3. Spearman's correlation coefficients for *LX*, *LY*, *LZ*, and *ldpc*.

	LX	*LY*	*LZ*
Correlation coefficient (r_s)	0.858	0.857	0.851
2-tailed $p - value$	<0.001	<0.001	<0.001

It can be assumed that *LX*, *LY* and *LZ* have a strong positive linear relationship with the *ldpc* since r_s was positive and near to the value of 1, with a significance value inferior to $\alpha = 0.01$. However, it is suspected that there exists multicollinearity between the independent variables *LX*, *LY* and *LZ*, which means that they could be related to one another, which indicates that if one of them has a high correlation with the dependent variable the others would have as well.

3.3 Multiple Linear Regression

The diamonds are characterized by their dimensions which are quantitative variables, but *cut*, *color*, *clarity*, and *table* are qualitative variables. By default, SPSS does not automatically consider qualitative variables as dummy variables, impairing its use in linear regression methods. Therefore, the qualitative variables were converted into dummy variables as an intermediate step. Meaning that they were coded into numerical data, enabling the accurate building of the model. The importance of each one of those variables to influence the diamonds price was assessed through multiple linear regression, using stepwise selection methods.

Table 4 shows the results of the multiple linear regression of the selected independent variables (see Table 1), taking into consideration their individual coding. The variable selection techniques considered were the enter method, forward selection, and backward elimination. The models were assessed according to their adjusted coefficient of determination (R_a^2), Akaike Information Criterion (AIC) along with analysis of the resulting residuals assumptions and Cook's distance.

After iteratively conducting changes to the sequentially created multiple linear regression models, a satisfactory solution was obtained. The final model (indicated in the last row of Table 4), obeyed every assumption and had an R_a^2 of 88.3%, thus successfully explained the dependent variable behavior (*ldpc*).

Table 4. Model regression summary.

Model	R_a^2	AIC	Residuals	Outliers	Cooks distance
Enter	0.877	−5549.771	✓	6	0.064
Backward/Forward	0.877	−5551.375	✓	6	0.069
Backward (Final model)	0.883	−5564.973	✓	0	0.055

Two variables were eliminated from the final model. The first one to be cut off from the equation was *table*, which every model identified as not significantly contributing to the model and was removed in the forward and backward methods. The elimination of this variable improved the AIC and maximum Cook's distance value (decreased), but outliers were encountered nonetheless. More iterations were conducted, and the outliers were identified and removed from the dataset. Additionally, the variable (LY) was removed based on the previous knowledge gathered from the exploratory analysis, in which a strong and positive correlation was identified amongst the quantitative variables, related to the size of the diamonds. This strong correlation led to suspected multicollinearity among them, which in turn destabilizes the model [2]. This is also perceived in the results of the previous models, where the coefficient for the LY variable exhibited a negative tendency, contrary to its expected tendency. Once again, this result is incoherent and thus, it is assumed that this variable contributes to the destabilization of the model.

The analysis of residuals assumptions was conducted. Figure 4 is a Normal Q-Q Plot to validate the normality assumption. Most of the values are according to the diagonal. Thus, this assumption is satisfied.

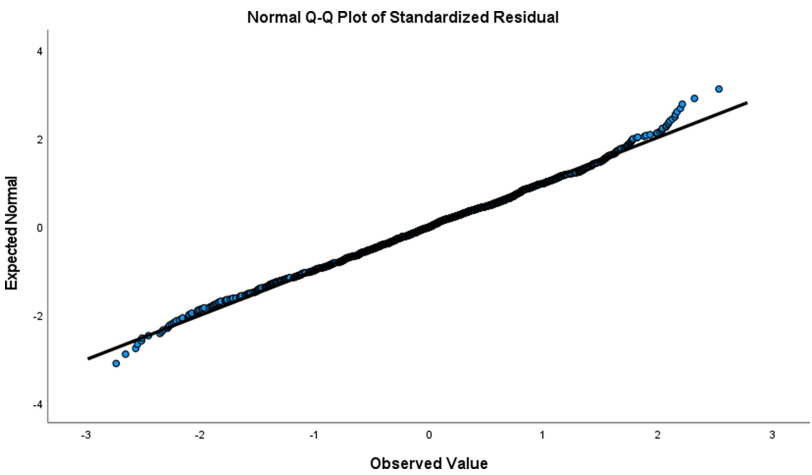

Fig. 4. Normal Q-Q Plot for the regression analysis residuals.

The homoscedasticity assumption was also validated by plotting the residuals (Fig. 5). It can be seen that does not exist any pattern and the values are randomly distributed.

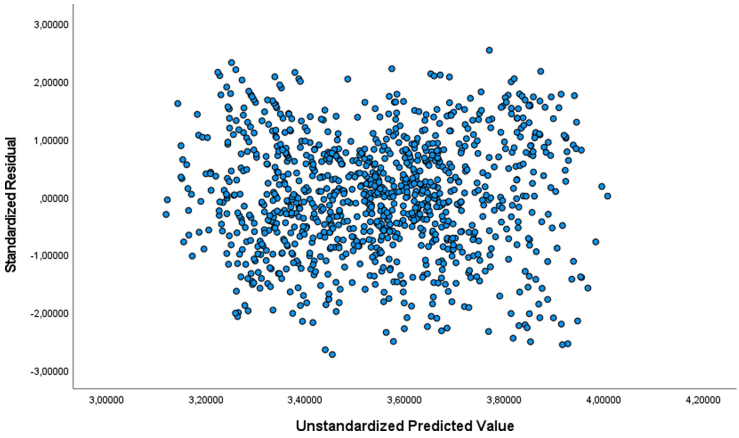

Fig. 5. Residuals plot.

In Fig. 6, it is presented the boxplot for residuals and no outliers are identified. Using the residuals descriptive statistics, the mean is equal to zero. Thereby, all the residuals assumptions were successfully validated.

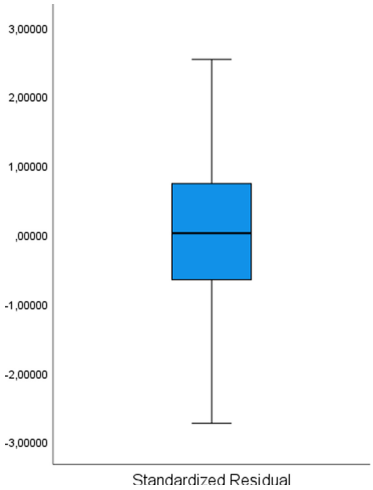

Fig. 6. Boxplot for residuals.

Having removed both *table* and *LY* variables along with the validation of all the assumptions of the final model, the backward and forward methods were both equally successful. The AIC value was the lowest reported one and the Cook's distance values did not surpass 0.06.

The coefficients and significance values of each variable are present in Table 5, except for the higher levels of the qualitative variables (ideal cut, colorless, and internally flawless), which are the reference levels. Therefore, the coefficient of the remaining qualitative variables is based upon a comparison against these reference levels.

Table 5. Coefficients of the final model.

Variable	Coefficient	P-value
Constant	2.051	<0.001
LX	1.629	<0.001
LZ	0.904	<0.001
Fair cut	−0.079	<0.001
Good cut	−0.032	<0.001
Very good cut	−0.019	0.001
Premium cut	−0.012	0.037
Near Colorless	−0.080	<0.001
Seriously flawed	−0.449	<0.001
Slightly included	−0.227	<0.001
Very slightly included	−0.128	<0.001
Very slightly included	−0.038	0.002

From a critical point of view, it is clear the importance of the *carat* variable. While the other variables did not behave as expected when the price was individually analyzed as the dependent variable, when *ldpc* is considered, the remaining variables acquired sense. Taking a closer look at the coefficients of the final model, exhibited in Table 5 tendencies can be identified.

The additive constant of the model is 2.051 and, in terms of the physical dimensions, it can also be pointed out that the length of the diamond (*LX*) weights more in this model than the depth (*LZ*), given the magnitude of the standardized coefficients. Regarding the qualitative variables, the negative coefficient is higher, the lower the corresponding level, and this is coherent with the relative importance of each of the levels. As an example, the *fair cut* is the most negatively affected level of the *cut* variable (−0.079), followed by the *good cut* (−0.032), then *very good cut* (−0.019), and finally, the *premium cut*, which is only affected by a coefficient of −0.012 in comparison with the *Ideal cut*. Moreover, it is worth noting that the lowest level of the cut variable (−0.079) is in the same range as the lowest *Color* level (−0.080) which means they negatively

affect the dependent variable within these limits: $[0, -0.080]$. On the other hand, the variable *Clarity*, which includes the remaining levels, affects the dependent variable in a more negative manner: $[0, -0.449]$.

4 Conclusions

In the current study, the analysis of influential factors on the diamond's price was carried out along with the proposal of an appropriate multiple linear regression model. To avoid high computational strain when processing data, a randomized sample was selected for analysis. In agreement with the established literature, the quantitative variables were transformed, which allowed a more accurate and balanced result. Moreover, it was necessary to transform the qualitative variables into dummy variables since it is a shortcoming of SPSS when the linear regression is performed.

It was possible to conclude that the variables *table* and *clarity* affect the dependent variable. Moreover, the quantitative variables have a strong positive linear relationship with the diamonds' price per carat. Stepwise selection models were conducted to evaluate which independent variables should be included in the model. As a result, the variable *table* was deemed not-significant by the selection models, and the *LY* was related to model destabilization. After removing them, the model was validated with an adjusted coefficient of determination of 88.3%. Besides, the final model was compliant with all the residuals' assumptions and encompassed all the remaining variables. The AIC value was reported as the lowest when compared to the other generated models and the tendencies experienced by the coefficients of the levels of the qualitative variables were coherent with the quality/price assumptions reviewed in the literature. The multiple linear regression helped define the importance of each set of variables in the final dependent variable. The variables *clarity* and length were identified as the most significant ones, in terms of qualitative and quantitative variables, respectively.

As far as future studies are concerned, the current study can be expanded upon by applying the Variance Inflation Factor (VIF), as a way of providing further insights into the multicollinearity issue that occurs among the quantitative variables. To address the identified constraints, Principal Components Analysis (PCA) can be applied to eliminate the multicollinearity effects. In addition, dividing the dataset into training and test sets can also be suggested as future work to evaluate the performance of the model to predict the diamonds' price.

References

1. Agrawal, S.: Diamond's price and atributes dataset (2007). https://www.kaggle.com/shivam2503/diamonds/metadata. Accessed 06 Oct 2020
2. Antunes, A.R., Braga, A.C.: Shiny app to predict agricultural tire dimensions. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12251, pp. 247–260. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58808-3_19

3. Bedoui, R., Guesmi, K., Kalai, S., Porcher, T.: Diamonds versus precious metals: what gleams most against USD exchange rates? Finance Res. Lett. **34**, 101253 (2020). https://doi.org/10.1016/j.frl.2019.08.001
4. Bundy, F.P., Hall, H.T., Strong, H.M., Wentorfjun., R.H.: Man-Made Diamonds. Nature **176**(4471), 51–55 (1955). https://doi.org/10.1038/176051a0
5. Busch, J.V., Dismukes, J.P.: Trends and market perspectives for CVD diamond. Diamond Related Mater. **3**(4), 295–302 (1994). https://doi.org/10.1016/0925-9635(94)90175-9
6. Cardoso, M.G., Chambel, L.: A valuation model for cut diamonds. Int. Trans. Oper. Res. (2005). https://doi.org/10.1111/j.1475-3995.2005.00516.x
7. Cavanaugh, J.E., Neath, A.A.: The Akaike information criterion: background, derivation, properties, application, interpretation, and refinements. WIREs Comput. Stat. **11**(3), e1460 (2019). https://doi.org/10.1002/wics.1460
8. Chatterjee, S., Simonoff, J.S.: Handbook of Regression Analysis. Wiley, Hoboken (2012). https://doi.org/10.1002/9781118532843
9. Chu, S.: Pricing the C's of Diamond Stones. J. Stat. Educ. **9**(2) (2001). https://doi.org/10.1080/10691898.2001.11910659
10. Falls, S.: Clarity, cut, and culture: The many meanings of diamonds (2014). https://doi.org/10.5860/choice.186333
11. Forthofer, R.N., Lee, E.S., Hernandez, M.: Linear Regression. In: Forthofer, R.N., Lee, E.S., Hernandez, M. (eds.) Biostatistics (Second Edition), pp. 349–386. Academic Press, San Diego, second edi edn. (2007). https://doi.org/10.1016/B978-0-12-369492-8.50018-2
12. Glatting, G., Kletting, P., Reske, S.N., Hohl, K., C, R.: Choosing the optimal fit function: comparison of the Akaike information criterion and the F-test. Med. Phys. **34**(11), 4285–4292 (2007). https://doi.org/10.1118/1.2794176
13. Harris, J.W.: Diamond. In: Encyclopedia of Geology, pp. 455–472. Elsevier (2021). https://doi.org/10.1016/B978-0-12-409548-9.12083-4
14. IBM Corp.: Ibm spss statistics 27 documentation (2020). https://www.ibm.com/support/pages/ibm-spss-statistics-27-documentation. Accessed 03 Feb 2021
15. Kimura, K., Waki, H.: Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. Optim. Methods Softw. **33**(3), 633–649 (2018). https://doi.org/10.1080/10556788.2017.1333611
16. Li, G., Rahim, M.Z., Pan, W., Wen, C., Ding, S.: The manufacturing and the application of polycrystalline diamond tools - a comprehensive review. J. Manuf. Processes **56**, 400–416 (2020). https://doi.org/10.1016/j.jmapro.2020.05.010
17. Mason, R.L., Gunst, R.F., Hess, J.L.: Statistical Design and Analysis of Experiments, 2nd edn. Wiley, New York (2003)
18. McKnight, P.E., Najab, J.: Mann-Whitney U Test. The Corsini Encyclopedia of Psychology (2010). https://doi.org/10.1002/9780470479216.corpsy0524
19. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. Int. Stat. Rev. **81** (2013). https://doi.org/10.1111/insr.12020_10
20. Montgomery, D.C.: Design and Analysis of Experiments, 8th edn. Wiley
21. Riffenburgh, R.H.: Statistical Prediction. In: Riffenburgh, R.H. (ed.) Statistics in Medicine, 2nd edn., pp. 125–136. Academic Press, Burlington (2006). https://doi.org/10.1016/B978-012088770-5/50047-2
22. Rutherford, A.: ANOVA and ANCOVA: A GLM Approach: Second Edition (2013). https://doi.org/10.1002/9781118491683
23. Salkind, N.J.: Statistics for people who (think they) hate statistics: Excel 2007 edition (2010)

24. Schwander, M., Partes, K.: A review of diamond synthesis by CVD processes. Diamond Related Mater. **20**(9), 1287–1301 (2011). https://doi.org/10.1016/j.diamond.2011.08.005
25. Schwertfeger, H., Fokin, A.A., Schreiner, P.R.: Diamonds are a chemist's best friend: diamondoid chemistry beyond adamantane. Angewandte Chemie Int. Edition **47**(6), 1022–1036 (2008). https://doi.org/10.1002/anie.200701684
26. Tillmann, W.: Trends and market perspectives for diamond tools in the construction industry. Int. J. Refractory Metals Hard Mater. **18**(6), 301–306 (2000). https://doi.org/10.1016/S0263-4368(00)00034-2
27. Verma, J.P.: Data analysis in management with SPSS software (2013). https://doi.org/10.1007/978-81-322-0786-3
28. Wilcox, R.R.: Understanding the practical advantages of modern ANOVA methods. J. Clin. Child Adolescent Psychol. (2002). https://doi.org/10.1207/S15374424JCCP3103_12
29. Wilcox, R.R.: 14 - MORE REGRESSION METHODS. In: Wilcox, R.R. (ed.) Applying Contemporary Statistical Techniques, pp. 517–555. Academic Press, Burlington (2003). https://doi.org/10.1016/B978-012751541-0/50035-3
30. Zhang, C.Y., Oh, S., Park, J.: State-of-the-Art Diamond Price Predictions using Neural Networks
31. Zimper, A.: On the welfare equivalence of asset markets and banking in Diamond Dybvig economies. Econ. Lett. **121**(3), 356–359 (2013). https://doi.org/10.1016/j.econlet.2013.09.023