Data Analytics Capstone Topic Approval Form

Student Name: Leng Yang

Student ID: 012298452

Capstone Project Name: Multivariate Linear Regression Analysis on the Sales Price of Cubic Zirconia

Project Topic: Exploring and determining factors that significantly impact the sales price of cubic zirconia.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: What attributes significantly influence the sales price of cubic zirconia?

Hypothesis:

Null Hypothesis: There is no significant association between any of 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.

Context:

Often, diamonds are used for engagement rings, but with how expensive the gemstone can be, it can be off-putting to potential customers. Not only that, but the controversy over blood diamonds also exists wherein diamonds are illegally mined or human rights abuses surrounding the working conditions of the miners (Baker, 2021). For these reasons, customers may look for choices, which leads to cubic zirconia, a popular lab-grown diamond alternative. The gemstone resembles a diamond in that it is naturally colorless and gives off a similar shine under natural light. The significant difference is the price, where cubic zirconia is much cheaper (Antolik, 2022).

Given this context and the dataset, this analysis aims to utilize linear regression to model and identify cubic zirconia attributes that significantly impact the prediction of final sales prices. Insights can be gained by exploring how each attribute contributes to the final sales price and distinguishing between higher and lower profitable stones based on these attributes. Such insights would aid the business in market research and forecasting.

Data:

The Co-learning Lounge authored and hosted the dataset on Kaggle (2021). The Co-learning Lounge consists of a community of learners that provides modified datasets for machine learning and deep learning purposes. Because the data is hosted on Kaggle, it is available to use for academic and research purposes.

The dataset contains 26,967 records and 11 variables. Below is a table summarizing the variables. All variables are independent, with the ‘price’ variable being the dependent variable.

Variable	Data Type
ID	Continuous
Carat	Continuous
Cut	Categorical
Color	Categorical
Clarity	Categorical
Depth	Continuous
Table	Continuous
Length (mm)	Continuous
Width (mm)	Continuous
Height (mm)	Continuous
Price	Continuous

Data Gathering:
The dataset will be downloaded off of Kaggle, where it is hosted. Afterward, the data will be cleaned, where duplicate records will be removed, and outliers will be replaced or removed. Once cleaned, the data will be explored, looking at any initial relationships and understanding the structure better. Lastly, the data will be preprocessed before modeling.

Data Analytics Tools and Techniques:
Exploratory analysis will be conducted to gain a better understanding of the structure of the dataset variables. Once completed, the data will be preprocessed before any data modeling. The modeling technique used in this analysis is multivariate linear regression. These analysis techniques will be performed using Python within a Jupyter Notebook environment.

Justification of Tools/Techniques:

Exploration of the data is justified because the information gathered gives the analyst a better understanding of the data. Data exploration will aid in identifying any initial insights and patterns. Linear regression analysis was chosen for this research due to its ease of understanding. It simplifies the association between the independent and dependent variables, thus making it easier to understand how each independent variable contributes. The statistical significance of the model can be understood by looking at the calculated F-statistic. Additionally, this ease of understanding allows for better presentation to a broader audience with different levels of knowledge.

Python is used for this analysis due to its simplicity and syntax. This simplicity allows for better readability as well as for code debugging. Another reason for Python is that there are numerous open-source libraries and packages that support the broad field of data analysis and science. A Jupyter notebook is used due to the iterative nature of the environment, allowing for ease of writing and testing code.

Project Outcomes:

This analysis aims to identify individual attributes that have a significant impact on price predictions. In doing so, insights into how much each attribute contributes to the sales price will also be generated. Furthermore, the linear regression model will be evaluated in its usefulness by determining whether the model is statistically significant via the F-statistic and the probability of the F-statistic. With these results, a useful model for cubic zirconia price prediction is expected to be developed.

Projected Project End Date: 11/30/2024

Sources:

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

Baker, A. (2021, June 28). *Blood Diamonds*. Time. https://time.com/blood-diamonds/

Co-learning Lounge. (2021, January 29). *Gemstone Price Prediction*. Kaggle. https://www.kaggle.com/datasets/colearninglounge/gemstone-price-prediction/data

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 11/5/2024

Reviewed by:

Comments: Click here to enter text.