FACULTY OF INDUSTRIAL ENGINEERING

GRADUATION PROJECT II

PRESENTED TO: **Prof. Dr. MUSTAFA NAFİZ DURU**

ADVISOR: **Arş. Gör. MELİSA ÇALIŞKAN DEMİR**

TEAM MEMBERS:

SAAD HUSSAIN

IBRAHIM QAIDI

MOAMAD KHALED OMAR

TOPIC:

"Predicting Diamond Prices: Regression Analysis and TOPSIS Method in Python"

01-06-2023

Table of Contents

List of Figures

Abstract

Accurately estimating diamond prices is essential in the diamond industry and consumer decision-making processes (*Diamond Prices: How to Calculate a Diamond's Value & Worth*, n.d.). This paper presents a novel approach that integrates regression analysis and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method for diamond price estimation. The objective is to rank diamonds based on attributes such as carat, cut, colour, clarity, and other relevant factors to determine their relative importance in price estimation (*Diamond Characteristics: The Four Cs*, n.d.). The methodology involves collecting a comprehensive dataset of diamonds, performing data preprocessing, and employing regression analysis to establish the relationships between attributes and prices (*Diamonds | Kaggle*, n.d.). The regression model is built using Python, enabling the estimation of diamond prices based on the identified features. Following regression analysis, the TOPSIS method is employed to rank the diamonds. The criteria weights, calculated using established methods such as expert opinions or statistical analysis, are assigned to each attribute (Podvezko, 2009). A decision matrix is constructed, and the positive and negative ideal solutions are identified. Using the normalized attributes, Euclidean distances are calculated to measure the proximity of each diamond to the ideal solutions(Zavadskas et al., 2016). The relative closeness of each diamond is determined by considering its distance to the negative ideal solution relative to the sum of its distances to both the positive and negative ideal solutions. The diamonds are then ranked based on their relative closeness values, providing an estimate of their prices. The integration of regression analysis enhances the accuracy and robustness of the price estimation process. It enables a deeper understanding of the relationships between attributes and prices, leading to more informed and reliable estimates (Sykes, n.d.). The implemented Python code allows for replication and customization of the methodology in future research and practical applications. The paper concludes with a discussion of the findings, implications for the diamond industry, and opportunities for further research. The validation of the method's accuracy and performance through comparisons between estimated and actual prices ensures the reliability and practicality of the proposed approach. In summary, the integration of regression analysis and the TOPSIS method offers a comprehensive framework for diamond price estimation. This approach provides valuable insights for industry professionals and consumers, facilitating informed decision-making and enhancing the overall understanding of diamond pricing dynamics.

Keywords:

Multicriteria Decision Making, TOPSIS, AHP, Regression Analysis, Normalization, Python, Diamond Pricing, Correlation, Model, Dataset, Python.

1 Problem Definition

Accurately estimating the price of diamonds is of paramount importance in the diamond industry and consumer decision-making processes. The pricing of diamonds is influenced by various attributes, including carat, cut, colour, clarity, and other relevant factors(*Diamond Quality Factors*, n.d.). However, determining the relative importance of these attributes and their impact on diamond prices can be a complex task. The diamond market is a complex and highly competitive industry, where accurate pricing plays a crucial role in facilitating fair transactions and informed decision-making.

The existing approaches for diamond price estimation often rely on simplistic methods that do not adequately consider the interplay of multiple attributes. This limitation hampers the accuracy and reliability of price estimations, leading to potential discrepancies between estimated and actual diamond values. Therefore, there is a need for a robust methodology that comprehensively incorporates the various attributes of diamonds to provide more precise and reliable price estimations.

The primary problem addressed in this report is the need for a reliable and efficient method to estimate diamond prices based on their features. While various pricing models and approaches exist, the integration of regression analysis and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method offers a comprehensive solution to this problem.

Regression analysis enables the identification and quantification of the relationships between independent variables (diamond features) and the dependent variable (price). By analysing the coefficients obtained from the regression models, it becomes possible to determine the magnitude and direction of these relationships, providing valuable insights into the impact of each diamond feature on its price(Bijaya et al., 2019).

Additionally, the TOPSIS method allows for a holistic evaluation of the diamonds in the dataset. By considering multiple criteria and assigning weights to each criterion, TOPSIS enables the ranking of the diamonds based on their overall desirability and similarity to an ideal diamond. This ranking aids buyers and sellers in making informed decisions by providing a comparative assessment of the diamond specimens.

Thus, the problem can be defined as the development of a robust methodology that combines regression analysis and the TOPSIS method to estimate diamond prices accurately. The objective is to provide a comprehensive understanding of the relationships between diamond features and prices, as well as a practical tool for ranking and evaluating diamond specimens.

2 Background of Multi-Criteria Decision Making (MCDM)

Multi-criteria decision making (MCDM) has its roots in decision theory, operations research, and management science. The field emerged as a response to the limitations of traditional decision-making approaches that relied on a single criterion, often leading to oversimplified and inadequate solutions. MCDM addresses the complexities inherent in decision problems by considering multiple criteria simultaneously, providing a more comprehensive and robust decision-making framework(Köksalan et al., 2011).

Historical Development: The foundations of MCDM can be traced back to the mid-20th century when scholars began recognizing the need for a systematic approach to deal with decision situations involving multiple objectives. In the 1960s, researchers like Charles L. Cochrane, Howard Raiffa, and R. Duncan Luce made significant contributions to the field of decision theory, laying the groundwork for multi-criteria decision making(*Multiple Criteria Decision Making | International Society on MCDM*, n.d.).

During the 1970s and 1980s, advancements in MCDM methods gained momentum with the introduction of influential techniques. The Analytic Hierarchy Process (AHP), developed by Thomas L. Saaty, gained popularity for its ability to handle complex decision hierarchies. Other methods, such as the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) and the Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE), emerged during this period(Saaty, 1990).

The 1990s witnessed further developments in MCDM, with researchers exploring new techniques and refining existing ones. Mathematical programming-based approaches, including goal programming, linear programming, and fuzzy set theory, were integrated into MCDM methods to handle real-world decision problems more effectively(Kildiene, 2014). Additionally, advancements in computing power and software tools facilitated the implementation and practical application of MCDM methods.

MCDM is built on several fundamental concepts and principles, including:

1. Multiple Criteria: MCDM recognizes that decision problems involve multiple criteria or objectives that must be considered simultaneously. These criteria can be quantitative or qualitative and may represent different dimensions of the decision problem, such as cost, time, quality, risk, and environmental impact.

2. Trade-Offs and Preferences: MCDM acknowledges that decision makers must make trade-offs between criteria because optimizing one criterion often comes at the expense of another. Understanding decision makers' preferences and their willingness to trade-off between criteria is essential in MCDM.

3. Decision Hierarchy: MCDM often employs a hierarchical structure to organize the decision problem. The hierarchy represents the relationships between criteria, sub-criteria, and alternatives, facilitating a structured analysis of the decision problem.

4. Weighting and Aggregation: MCDM methods involve assigning weights to criteria to reflect their relative importance or preferences. Aggregation techniques are used to

combine the performance of alternatives across criteria and generate an overall evaluation.

5. Sensitivity Analysis: MCDM recognizes the importance of assessing the sensitivity of the decision outcomes to changes in criteria weights or alternative performance. Sensitivity analysis helps evaluate the robustness of the results and provides insights into the stability of the decision recommendations.

MCDM has had a significant impact across various fields and industries. Its applications are widespread and diverse, ranging from business and engineering to healthcare and environmental management. MCDM methods have been used in project selection, portfolio optimization, supplier evaluation, risk assessment, resource allocation, policy prioritization, and many other decision-making contexts(Jahan & Edwards, 2013).

The integration of MCDM into decision-making processes has resulted in more informed and robust decisions. By considering multiple criteria and accommodating stakeholders' preferences, MCDM has enhanced transparency, objectivity, and accountability in decision-making processes.

Future Directions: As the complexity of decision problems continues to grow, MCDM remains an active area of research and development. New techniques and hybrid approaches are constantly being explored to handle evolving challenges. Incorporating uncertainties, incorporating machine learning and artificial intelligence techniques, and addressing dynamic decision contexts are some areas where MCDM is evolving(Kildiene, 2014).

In conclusion, multi-criteria decision making has emerged as a powerful framework to address complex decision problems involving multiple objectives. With its foundations in decision theory and operations research, MCDM provides decision makers with systematic approaches and tools to evaluate alternatives, balance conflicting criteria, and arrive at informed decisions across a wide range of domains(Kumar et al., 2017). The ongoing advancements and applications of MCDM ensure its continued relevance and impact in the face of ever-increasing complexity in decision-making scenarios.

3 Multi-Criteria Decision Making: Balancing Complex Choices

Introduction: Multi-criteria decision making (MCDM) is a powerful methodology used to address complex decision-making scenarios that involve multiple conflicting objectives or criteria. In today's interconnected and data-driven world, decision makers often face situations where a single criterion is insufficient to capture the full complexity of a problem. MCDM provides a systematic framework for evaluating alternatives and selecting the most favourable option based on a range of criteria, leading to more informed and robust decisions(*Multi-Criteria Decision Analysis (MCDA/MCDM) | 1000minds*, n.d.).

1. Understanding MCDM: Multi-criteria decision making is an approach that considers multiple decision criteria simultaneously to evaluate alternatives. It recognizes that

decisions involve trade-offs and that no single criterion can fully capture the complexity of a decision problem. MCDM methods help decision makers analyze and weigh the importance of different criteria to arrive at an optimal or satisfactory decision.

2. Types of MCDM Methods: There are several MCDM methods available, each offering its unique approach and mathematical foundation. Some widely used methods include the Analytic Hierarchy Process (AHP), the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), the Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE), and the Weighted Sum Model (WSM). These methods incorporate various techniques such as pairwise comparisons, ranking, and aggregation to facilitate decision making.

3. Steps in the MCDM Process: The MCDM process typically involves the following steps: a. Problem Identification: Clearly define the decision problem and establish the criteria and alternatives that will be considered. b. Criteria Selection and Weighting: Identify the relevant decision criteria and assign weights to reflect their relative importance. c. Data Collection: Gather data and information on the alternatives and their performance with respect to the criteria. d. Evaluation and Analysis: Apply the chosen MCDM method to evaluate the alternatives and generate rankings or scores based on the criteria. e. Sensitivity Analysis and Interpretation: Assess the robustness of the results and interpret the findings in the context of the decision problem.

4. Advantages of MCDM: MCDM offers several advantages over traditional decision-making approaches: a. Comprehensive Evaluation: MCDM enables decision makers to consider multiple criteria simultaneously, leading to a more comprehensive evaluation of alternatives. b. Transparency and Consistency: The systematic nature of MCDM methods enhances transparency and provides a clear rationale for decision outcomes. c. Flexibility: MCDM methods can accommodate diverse decision contexts and criteria, allowing decision makers to customize the analysis to their specific needs. d. Consideration of Stakeholder Perspectives: MCDM facilitates the integration of multiple stakeholders' perspectives by incorporating their criteria and preferences into the decision-making process.

5. Applications of MCDM: MCDM has found applications in various fields, including business, engineering, environmental management, healthcare, and public policy. Examples include project selection, supplier evaluation, site selection, risk assessment, resource allocation, and policy prioritization. MCDM methods provide valuable insights and support decision-making processes in these complex domains.

Multi-criteria decision making is a valuable tool for tackling complex decision problems that involve multiple criteria. By considering multiple objectives and incorporating various decision factors, MCDM methods enable decision makers to make more informed and balanced choices. As our world becomes increasingly complex, MCDM continues to play a vital role in facilitating effective decision making across a wide range of disciplines and industries(Jahan & Edwards, 2013).

4 TOPSIS

The Technique for Order of Preference by Similarity to Ideal Solution, or TOPSIS for short, is a well-known MCDM strategy that ranks possible courses of action based on how closely they resemble the ideal solutions, both positive and negative. The positive ideal solution is the best response that can possibly be given, whereas the negative ideal solution is the worst answer that can possibly be given(Pal et al., 2013).

TOPSIS is a straightforward method of organizing that was developed by Hwang and Yoon in the year 1981(Bhutia & Phipon, 2012). By default, TOPSIS will search for solutions that are optimal in terms of how close they are to the positive ideal while also being optimal in terms of how far they are from the negative ideal before selecting one of those solutions. The strategy that follows the negative ideal seeks to maximize costs while mitigating benefits; in contrast, the approach that follows the positive ideal seeks to maximize benefits while minimizing costs. TOPSIS makes significant use of attribute data, rates choices using a cardinal scale, and does not require that attribute values be exclusive of one another in order to function correctly. (Chen and Hwang, 1992; Yoon & Hwang, 1995). In order for this technique to function, the quantities of the attributes must be integers, they must consistently increase or decrease, and they must have units that are equivalent. Figure 2 illustrates the step-by-step procedure that Hwang and Yoon (1981) developed for implementing TOPSIS. Adjustments are made to the decision matrix, which are created following the formation of the initial decision matrix. This is where the procedure starts. Secondly, a weighted normalized option matrix is constructed. Then, determining the positive and negative optimum solutions is carried out, and finally, the separation values for each of the potential courses of action is calculated. After everything else

is said and done, the proportionate closeness score is computed. The various possibilities (or possible candidates) can be ranked according to how close they are to the optimal answer.

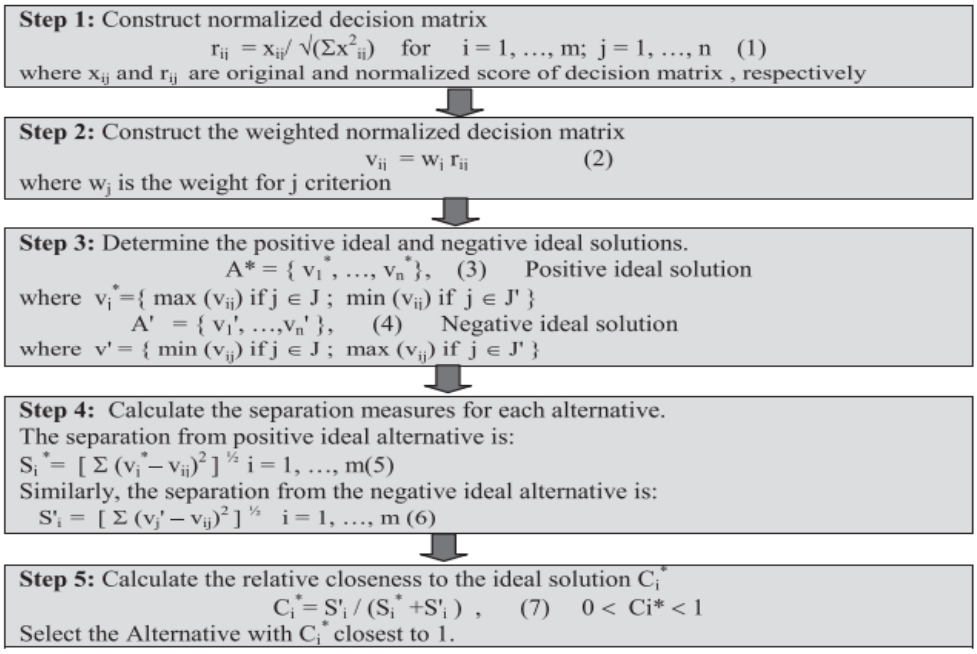

Figure 1: Stepwise guide to TOPSIS Method.

5 Regression Analysis

Regression analysis is a statistical method used to examine and quantify the relationship between a dependent variable and one or more independent variables. It aims to understand how changes in the independent variables influence the dependent variable(*Regression Analysis - Rudolf J. Freund, William J. Wilson, Ping Sa - Google Books*, n.d.). In the context of diamond pricing, regression analysis helps determine the impact of various diamond features on the price of a diamond specimen.

The dependent variable, in this case, is the price of the diamond, which is influenced by multiple independent variables such as carat weight, cut quality, color grade, clarity grade, and dimensions. Regression analysis enables us to estimate the effect of each independent variable on the dependent variable and provides insights into the direction and magnitude of these effects(Samprit Chatterjee, n.d.).

There are different types of regression analysis, but the most commonly used method is multiple linear regression. In multiple linear regression, the relationship between the dependent variable (price) and several independent variables (diamond features) is modelled through a linear equation. The equation takes the form:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$$

Where:

- ***Y*** represents the dependent variable (price),
- ***X1, X2, ..., Xn*** are the independent variables (diamond features),
- ***β0, β1, β2, ..., βn*** are the coefficients (also known as slopes) that represent the impact of each independent variable,
- ε is the error term, accounting for the variability that cannot be explained by the independent variables.

The regression analysis process involves estimating the coefficients (slopes) that minimize the sum of squared differences between the predicted and actual values of the dependent variable. This estimation is typically done using statistical techniques such as ordinary least squares (OLS) or maximum likelihood estimation (MLE).

Once the coefficients are determined, they provide valuable information about the relationship between the independent variables and the dependent variable. Positive coefficients indicate a positive relationship, meaning that an increase in the value of the independent variable leads to an increase in the dependent variable (price). Conversely, negative coefficients signify an inverse relationship.

By analysing the coefficients, one can assess the relative importance of each independent variable in determining the diamond price. The coefficients also enable price predictions for new diamond specimens based on their features, thereby aiding decision-making processes in the diamond industry.

Regression analysis serves as a powerful tool for understanding the factors that influence diamond prices, providing valuable insights into the pricing dynamics and facilitating more accurate price estimation.

6 Literature Review

The purpose of this literature analysis was to find papers from reputable publications that offer the most useful insights into the TOPSIS methodology for academics and practitioners tackling real-world problems. With this goal in mind, we combed through the titles, abstracts, and terms of academic articles looking for references to TOPSIS. Elsevier, Springer, Taylor & Francis, Emerald, John Wiley, IEE Explore, and EBSCO were the primary sources we sought out, as they house the majority of the top publications in operation research and management. Therefore, papers presented at conferences, theses and dissertations at all levels, manuals, and research papers that were never published were left out of the literature survey.

6.1 Application Domain:

Regression analysis is a versatile statistical technique that finds widespread application in numerous domains. In economics and finance, it is employed to assess the effects of variables like interest rates, inflation, income, and government policies on economic indicators such as GDP, investment, and consumption. Social sciences benefit from regression analysis by studying connections between variables like education level, income, demographics, and social outcomes like crime rates, health outcomes, and academic achievement(Marzouk & Sabbah, 2021). Market research utilizes regression analysis to examine consumer behaviour, predict preferences, and understand the impact of marketing variables on sales, market share, and customer satisfaction. In healthcare and medicine, regression analysis aids in evaluating the relationship between patient characteristics, treatments, and health outcomes, enabling the prediction of disease outcomes and identification of risk factors(Hansen & Devlin, 2019). Environmental science utilizes regression analysis to explore the connections between environmental factors, such as pollution levels and climate variables, and their effects on ecological outcomes. Despite its diverse applications, regression analysis has limitations, including assumptions of linearity, multicollinearity, sensitivity to outliers, and the need for cautious interpretation regarding causality(Pal et al., 2013). However, when used appropriately, regression analysis remains a valuable tool in understanding and modelling relationships in various fields.

TOPSIS is widely used in the building industry. As a result, engineers are frequently required to weigh competing priorities when making decisions. Engineers can use TOPSIS to find the optimal solution that meets these requirements(Sarkar, 2010). Using factors like strength, rigidity, density, and expense, TOPSIS can help choose the optimal substance for a task.

TOPSIS also has significant potential in the fields of business and administration. Profitability, market dominance, client happiness, and staff involvement are just some of the many factors that go into making business decisions. Business plans, product ideas, and funding possibilities can all be prioritized with TOPSIS's help(Zavadskas et al., 2016). TOPSIS can be used to determine the optimal spot for a new retail outlet based on factors like population density, level of rivalry, and ease of entry.

The social sciences have also benefited from TOPSIS's use. For instance, TOPSIS can be used to rate potential employees according to factors like their schooling, work history, and character characteristics. TOPSIS can also be used in medicine to rate the efficacy, safety, and expense of potential treatments(Marzouk & Sabbah, 2021).

TOPSIS is also useful for environmental control. Using factors like environmental influence, societal acceptability, and fiscal viability, TOPSIS can help choose the most effective environmental strategy or initiative(Zavadskas et al., 2016). Using factors like energy production, environmental effect, and expense, TOPSIS can be used to choose the most beneficial green energy initiative.

Finally, TOPSIS is a multi-criteria decision-making approach useful in many contexts. It's useful for making decisions in difficult circumstances because it ranks options according to how close they are to the ideal solution and how different they are from the negative-ideal solution. Decision-makers in any field, be it engineering, business, the social sciences, or

environmental management, can benefit from using TOPSIS to determine which option best suits their needs.

The TOPSIS technique has many practical applications, providing compelling reasons to classify its uses according to broad subject areas and narrower sub-disciplines. Studies involving actual applications typically incorporate case studies, instructive instances, and/or personal experiences. Two hundred and sixty-six papers were divided into nine groups according to their commonalities and distinctions. These groups were Supply Chain Management and Logistics (1), Design, Engineering, and Manufacturing Systems (2), Marketing and Business Management (3), Safety, Health, and Environmental Management (4), Personnel Management (5), the management of energy (6), the chemical engineering field (7), the management of water resources (8), and other topics (9). Where multiple options existed for a given paper, the one that best served the paper's stated purpose was chosen. This guarantees that our system of organization is free of any duplications. Publications in Medical Sciences, Food and Agriculture, Schooling, Design, Politics, and Sports are included in the final section. More than half of all published apps fall into just two broad categories: "Supply Chain Management and Logistics" and "Design, Engineering, and Manufacturing Systems." The fields of Chemical Engineering and Water Resources Management have received scant attention in terms of software development.

Areas	*N*	%
Supply Chain Management and Logistics	74	27.5
Design, Engineering and Manufacturing Systems	62	23
Business and Marketing Management	33	12.3
Health, Safety and Environment Management	28	10.4
Human Resources Management	24	8.9
Energy Management	14	5.2
Chemical Engineering	7	2.6
Water Resources Management	7	2.6
Other topics	20	7.4
Total	269	100

Figure 2 Distribution of Literature on TOPSIS by topic

6.2 Limitations

Since its beginning in the year 2000 by Chen, the FTOPSIS technique has been hampered by a variety of constraints, issues, and challenges, which we will describe in this article. FTOPSIS begins the process of data collection by first analyzing the numerical numbers of characteristics. A key component of FTOPSIS is the calculation of distance measures between each option and the fuzzy PIS and fuzzy NIS. Due to the fact that characteristics are presented,

this approach is classified as Multi-Attribute Decision Making (MADM or MCDM).

No.	Characteristics	TOPSIS
1	Category	Cardinal information, information on attributes, MADM
2	Core process	The distances from PIS and NIS(cardinal absolute measurement)
3	Attribute	Given
4	Weight elicitation	Given
5	Consistency check	None
6	Reliability check	None
7	No. of attributes accommodated	Many more
8	No. of alternatives accommodated	Many more
9	Others	Compensatory operation

Figure 3: Summary of characteristics of TOPSIS method.

Regression analysis, while widely employed and versatile, is subject to several limitations. One limitation is the assumption of linearity between independent and dependent variables(Douglas C. Montgomery, Elizabeth A. Peck, n.d.). In reality, relationships can be nonlinear, and relying on a linear model may yield inaccurate predictions or biased estimates. Multicollinearity is another challenge, arising when independent variables are highly correlated, making it difficult to interpret coefficients and leading to unstable estimates(Liang & Zeger, 2003). Outliers, influential observations that deviate significantly from the norm, can exert undue influence on estimated coefficients, resulting in model misspecification and reduced prediction accuracy. Assumptions regarding the distribution and variance of residuals, such as normality and homoscedasticity, may be violated, compromising the validity of regression results. It is important to note that regression analysis reveals associations but does not establish causality, necessitating caution in interpreting the findings(Norman R. Draper, n.d.). Finally, the extrapolation of regression models beyond the observed data range may produce unreliable predictions. Despite these limitations, researchers can mitigate these challenges through diagnostic techniques and robust modelling approaches, ensuring the validity and reliability of regression analysis for understanding and modelling variable relationships.

7 Methodology

The objective of this research is to estimate the price of a diamond specimen using both Regression Analysis and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) analysis. The methodology involves several steps, including data collection, data pre-processing, criteria identification, regression analysis, and TOPSIS analysis. The updated methodology with steps for regression analysis is as follows:

1. Data Collection:
    - Obtain a dataset in CSV format that contains information about diamonds, including attributes such as carat, cut, color, clarity, and price.
2. Data Pre-processing:
    - Load the CSV data into a pandas DataFrame in Python.
    - Perform any necessary data cleaning, such as handling missing values or outliers.
    - Convert categorical variables (e.g., cut, color, clarity) into numerical representations, if required, using label encoding.
3. Criteria Identification:
    - Identify the relevant criteria for estimating the price of a diamond. These criteria may include carat, cut, color, clarity, and any other attributes deemed significant.
4. Regression Analysis:
    - Find the correlation matrix to understand the relationships between the independent variables (diamond features) and the dependent variable (price).
    - Plot regression lines or scatter plots to visualize the relationships between the independent variables and the dependent variable.
    - Build a regression model to train the data and estimate the price of a diamond specimen. This involves selecting an appropriate regression algorithm (e.g., linear regression, multiple regression) and evaluating the model's performance using metrics such as R-squared, mean squared error, or root mean squared error.
    - Implement a pipeline to streamline the steps involved in the regression analysis, including data preprocessing, model training, and evaluation.
5. TOPSIS Analysis:
    - Normalize the numerical attributes (e.g., carat) to eliminate differences in scales using techniques like min-max scaling or z-score normalization.
    - If necessary, normalize the categorical attributes by applying appropriate transformations or encoding techniques.
    - Assign weights to each criterion based on their relative importance in estimating the price of a diamond.
    - Construct the decision matrix by multiplying each element of the normalized data by its corresponding criterion weight.
    - Determine the positive ideal solution and negative ideal solution based on the constructed decision matrix.

- Calculate the Euclidean distances between each diamond specimen and the ideal solutions.
- Compute the relative closeness for each diamond specimen based on the calculated distances.
- Rank the diamond specimens based on their relative closeness values to estimate the price.

6. Interpretation and Validation:
    - Interpret the results of both the regression analysis and TOPSIS analysis to understand the relative importance of different criteria in estimating diamond prices.
    - Validate the regression analysis results by comparing the estimated prices with the actual prices of the diamonds. Evaluate the accuracy and performance of the regression model.
    - Validate the TOPSIS analysis results by analyzing the rankings and comparing them with the regression analysis results.
7. Discussion and Conclusion:
    - Discuss the findings of both the regression analysis and TOPSIS analysis and their implications for estimating diamond prices.
    - Summarize the methodology, limitations, and future directions for further improvement.

By combining Regression Analysis and TOPSIS analysis, this methodology provides a comprehensive approach to estimating diamond prices, leveraging the strengths of both techniques and ensuring robust results.

8 Implementation: Python Code for TOPSIS Analysis

Step 1:

Downloading the appropriate libraries. The use of NumPy, pandas, Matplotlib, and Seaborn in data analysis and visualization tasks has become prevalent in the Python ecosystem. These libraries offer powerful capabilities for handling data, performing numerical computations, and creating insightful visualizations. Moreover, numerous modules for ***'sklearn'*** have been imported. ***'sklearn'*** specializes in machine learning and statistical data analysis. Modules such as '***LinerRegression'***, '***LabelEncoder'*** will be described in further detail below.

```
In [125]:
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib as mpl
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.pipeline import Pipeline
from sklearn.tree import DecisionTreeRegressor

In [126]:
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import cross_val_score
from sklearn.metrics import mean_squared_error
from sklearn import metrics

In [127]:
pip install xgboost

Requirement already satisfied: xgboost in c:\users\saads\anaconda3\lib\site-packages (1.7.5)
Requirement already satisfied: scipy in c:\users\saads\anaconda3\lib\site-packages (from xgboost) (1.10.0)
Requirement already satisfied: numpy in c:\users\saads\anaconda3\lib\site-packages (from xgboost) (1.23.5)
Note: you may need to restart the kernel to use updated packages.

In [128]:
from xgboost import XGBRegressor
```

Figure 4: Importing Libraries.

Step 2:

Uploading the .csv file (comma separated values) containing information about various attributes regarding diamonds and displaying some values. The file was provided by Shivam Agarwal on Kaggle as an open data source by the name of 'Diamonds'.

```
In [234…
df = pd.read_csv(r'C:\Users\saads\OneDrive\Desktop\Diamonds - Dataset\diamonds.csv')
df.head()
```

Out[234]:

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Figure 5: Preprocessed Data (top 5 rows displayed).

In order to better understand the features of a diamond, the image below helps visualize features like ‘depth’ and ‘table’. ***Total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)*** The depth of the diamond is its height (in millimetres) measured from the culet (bottom tip) to the table (flat, top surface). A diamond's table refers to the flat facet of the diamond seen when the stone is face up. The main purpose of a diamond table is to refract entering light rays and allow reflected light rays from within the diamond to meet the observer’s eye. The ideal table cut diamond will give the diamond stunning fire and brilliance.

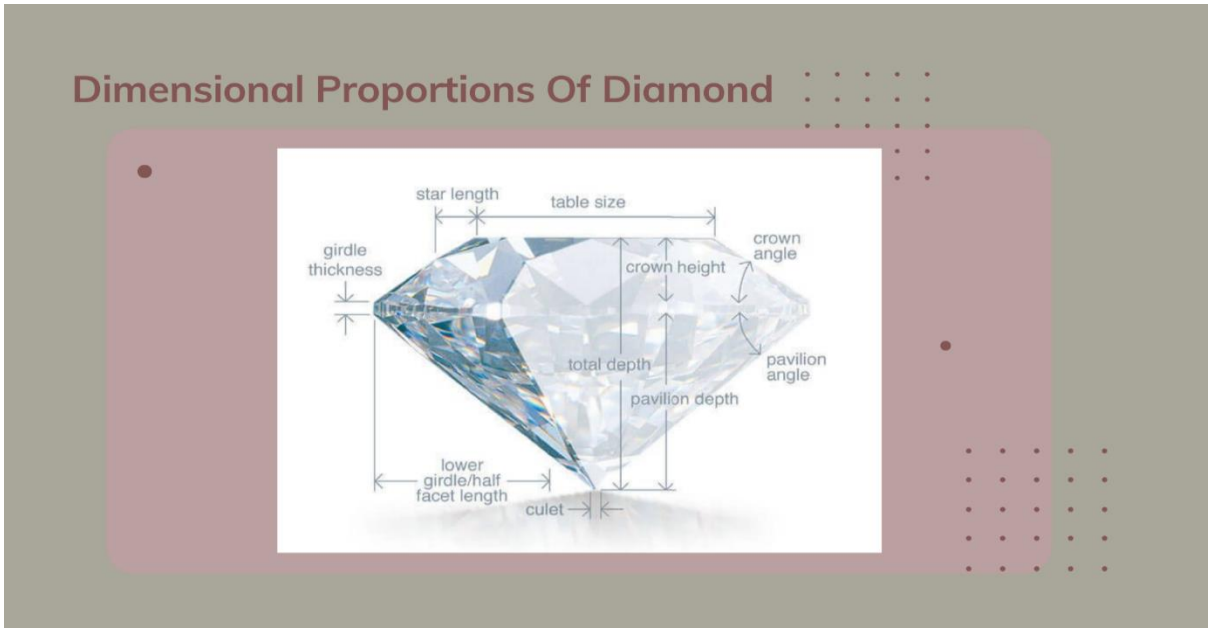

Figure 6: Features of a Diamond Visualized.

Step 3:

Replacing non-numeric values with numeric values. Diamond Cut, Clarity, and Colour are considered to have ordinal relations because they possess inherent ordering and gradations that impact the quality and value of a diamond. Due to these ordinal relationships in the dataset being used, there is the possibility of simply replacing ‘object’ values in the data with ‘integer’ or ‘float’ which increases in value as the quality of a feature increases; this approach is called Integer Encoding. However, using Label Encoding is a much better approach as it simplifies the code and makes it more efficient. The result for both methods is identical.

1. Diamond Cut: The cut of a diamond refers to the quality of its proportions, symmetry, and polish, which directly influence its brilliance and sparkle. Cut grades, such as "Excellent," "Very Good," "Good," "Fair," and "Poor," represent a scale that reflects the quality and desirability of the diamond's cut. These grades are assigned based on expert assessments of how well the diamond interacts with light and how effectively it reflects and refracts light to create its signature sparkle. The order of these grades signifies the increasing level of cut quality, making it an ordinal attribute.
2. Diamond Clarity: Clarity assesses the presence of internal or external characteristics, known as inclusions and blemishes, respectively, within a diamond. Clarity grades, such as "Flawless" (FL), "Internally Flawless" (IF), "Very Very Slightly Included"

(VVS1/VVS2), "Very Slightly Included" (VS1/VS2), "Slightly Included" (SI1/SI2), and "Included" (I1/I2/I3), reflect the relative visibility and extent of these characteristics. The grading scale implies a hierarchical order where the clarity grade improves from I1 (lowest) to FL (highest), representing the increasing rarity and desirability of diamonds with fewer inclusions and blemishes.

3. Diamond Colour: The colour of a diamond pertains to its lack of colour, or rather, the presence of subtle tints or hues. The Gemmological Institute of America (GIA) grades diamond colour on a scale from D (colourless) to Z (light yellow or brown). The scale progresses in alphabetical order, indicating the increasing presence of colour. Diamonds with higher color grades (e.g., D, E) are considered more desirable and valuable due to their rarity and closer approximation to a pure, colourless state.

In [145]:
```
# Make copy to avoid changing original data
label_data = data.copy()

# Apply label encoder to each column with categorical data
label_encoder = LabelEncoder()
for col in object_cols:
    label_data[col] = label_encoder.fit_transform(label_data[col])
label_data.head()
```

Out[145]:

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	2	1	3	61.5	55.0	326	3.95	3.98	2.43
1	0.21	3	1	2	59.8	61.0	326	3.89	3.84	2.31
2	0.23	1	1	4	56.9	65.0	327	4.05	4.07	2.31
3	0.29	3	5	5	62.4	58.0	334	4.20	4.23	2.63
4	0.31	1	6	3	63.3	58.0	335	4.34	4.35	2.75

Figure 7: Label Encoding (converting categorical values to numeric values).

These ordinal relationships in diamond attributes allow for meaningful comparisons and rankings when evaluating and determining the quality and value of diamonds. They provide a basis for distinguishing between diamonds based on their respective levels of cut precision, clarity characteristics, and color grades, aiding both industry professionals and consumers in making informed decisions when purchasing or valuing diamonds.

Step 4:

Converting every variable to type float for ease in calculations. Since values for most attributes take decimal numbers, float is the most appropriate type for our DataFrame.

In [239…

```
#Converting every non-float variable to float
df['cut'] = df['cut'].astype(float)
df['color'] = df['color'].astype(float)
df['clarity'] = df['clarity'].astype(float)
df['price'] = df['price'].astype(float)
print(df.dtypes)
```

```
carat      float64
cut        float64
color      float64
clarity    float64
depth      float64
table      float64
price      float64
x          float64
y          float64
z          float64
dtype: object
```

Figure 8: Converting all datatypes to float.

Step 5:

Depicting the spread of data via Pair-plots. A pair plot is a plot of subplots where each subplot represents a bivariate distribution of two variables in the given dataset. Pair-plots will help visualize every feature's relationship with the other variables. We use ***'ax = sns.pairplot()'*** to visualize the data.

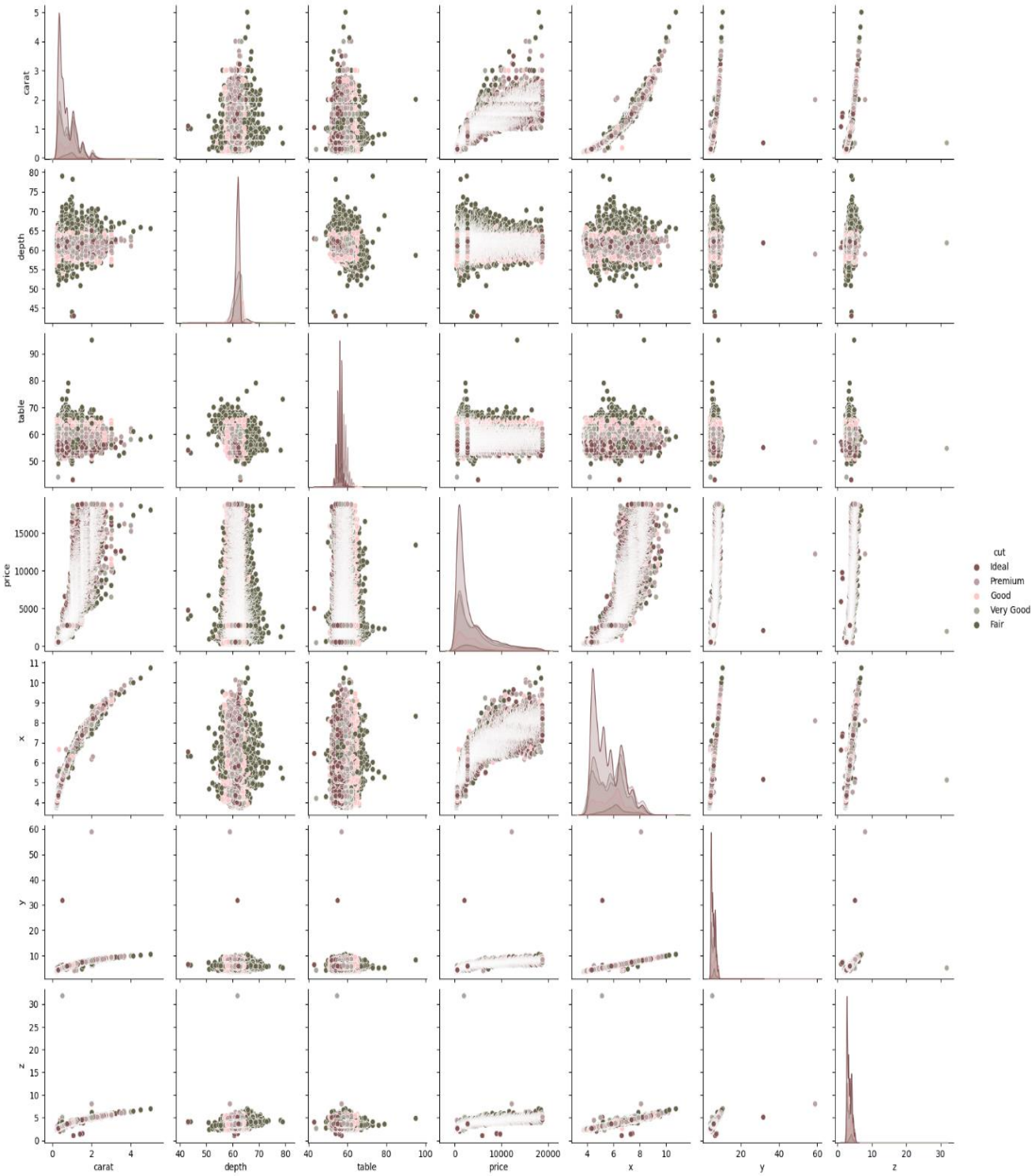

Figure 9: Pair-plot of uncleaned data.

Step 6:

In order to predict the price, it is separated from the dataset and defined as the only dependent variable. All other independent variables have an impact on the price of a diamond, in order to understand how each variable effects, the price, depicting regression lines between features and price is a good way to visually understand the data. As observed below, most features have a linear relationship with price.

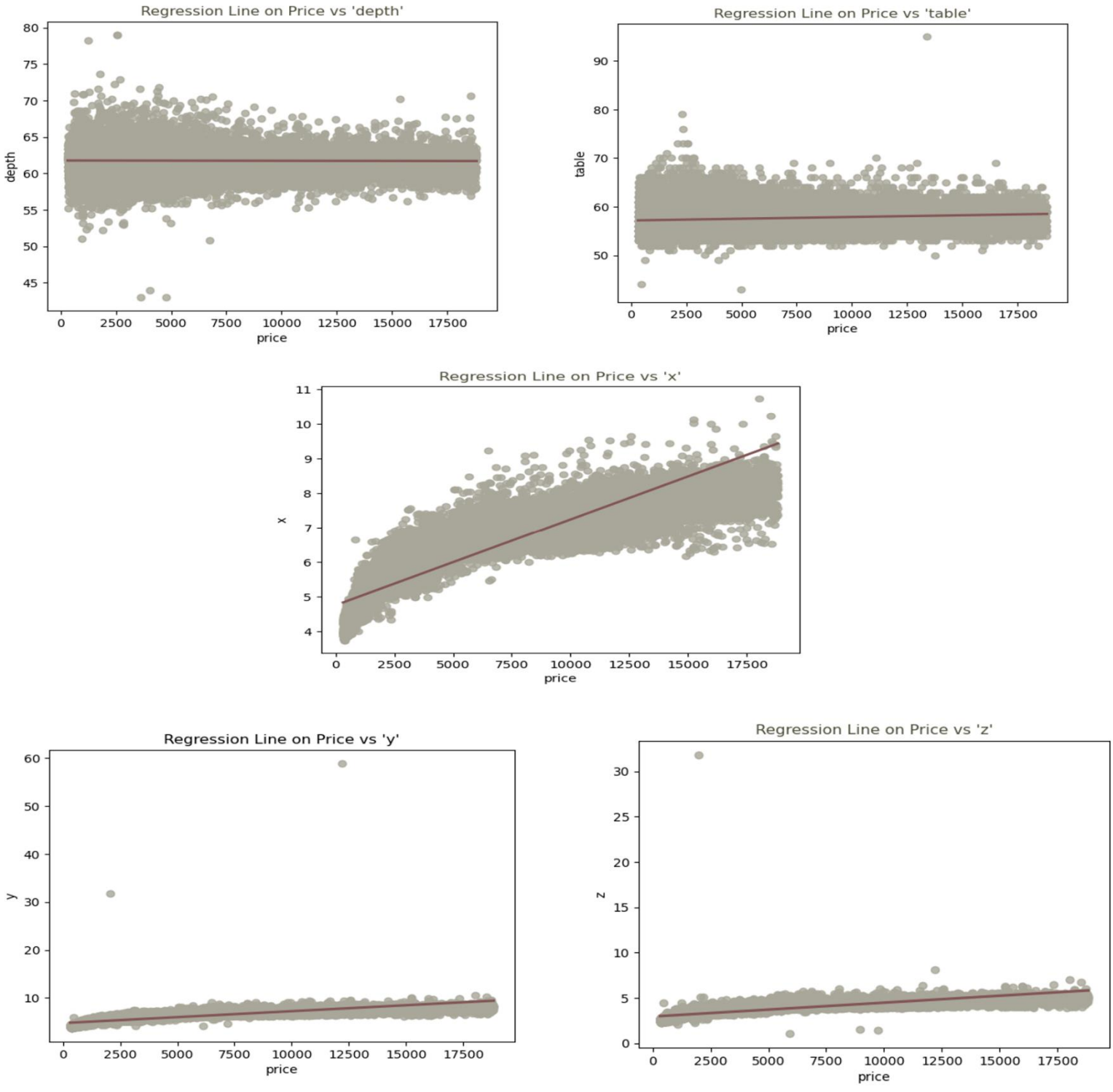

Figure 10: Linear Regression lines (price against features)

Step 7:

As observed in the scatter plots above, there are some outliers in certain variables. If a regression analysis model is run on this data, such points will skew our results. Therefore, these outliers are removed. To achieve this the range in which majority of the scatter points are observed is kept, and any point that does not lie within this range will be removed.

In [139]:
```
#Dropping the outliers.
data = data[(data["depth"]<75)&(data["depth"]>45)]
data = data[(data["table"]<80)&(data["table"]>40)]
data = data[(data["x"]<30)]
data = data[(data["y"]<30)]
data = data[(data["z"]<30)&(data["z"]>2)]
data.shape
```

Out[139]:
```
(53907, 10)
```

Figure 11: Removing outliers.

Once the outliers are removed, it concludes our preprocessing stage. The cleaned or preprocessed data is depicted below via Pair-plots.

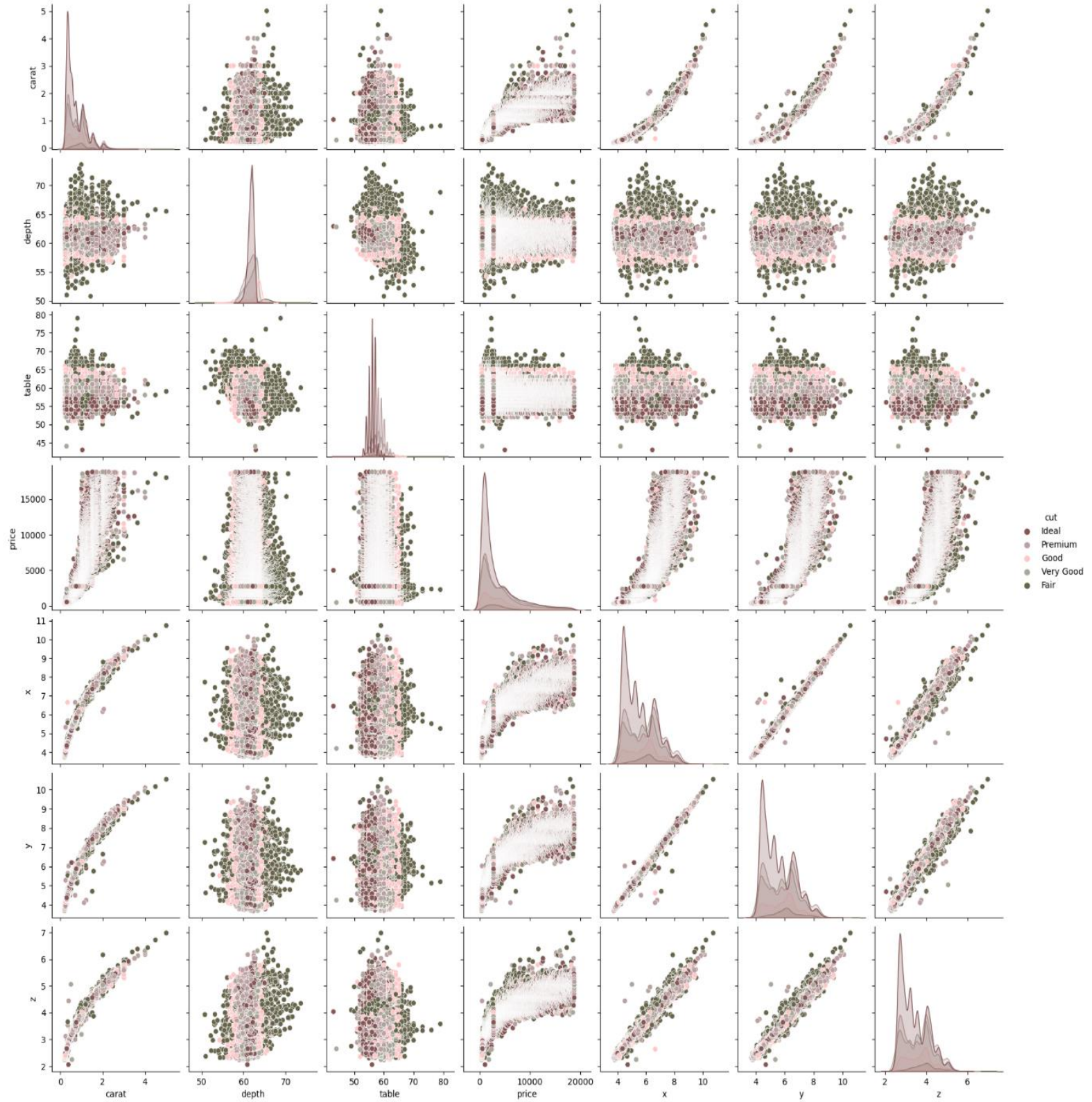

Figure 12: Pair-plots of cleaned data.

Step 8:

At this stage, data analysis techniques can be applied to further understand the relationship between each variable. The most efficient way to visualize this interrelationship is via a correlation matrix. A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship. Correlation is calculated by using ***'.corr'*** function. It is visualized by a heatmap below:

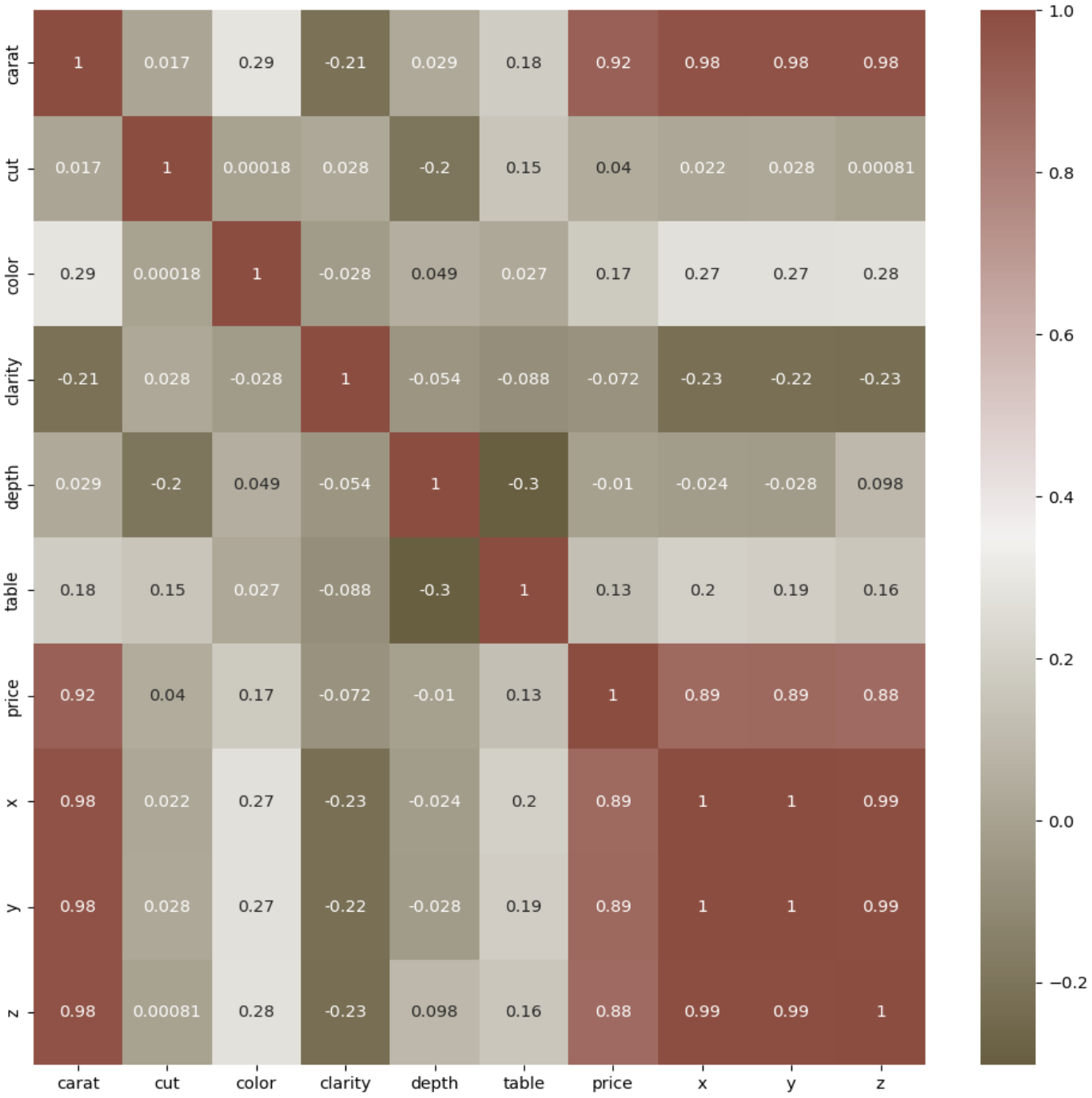

Figure 13: Correlation Matrix Heatmap.

Step 9:

Whenever machine learning is employed, it is recommended to build a training model and a test model. This allows the machine learning model to be trained using a portion of the dataset, while the remaining portion of the data is used to test the model's results. In the case of our dataset, the model will be trained to predict the price of a diamond, with the dependent variable ***'y'*** representing the price and the independent variables representing the features. This process is executed by the following code:

In [149]:
```
#MODEL BUILDING
# Assigning the featurs as X and trarget as y
X= label_data.drop(["price"],axis =1)
y= label_data["price"]
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.25, random_state=7)
```

Figure 14: Model Building - Splitting dataset to test and train.

Step 10:

The process of constructing a data preprocessing pipeline that includes the StandardScaler transformation and a regression model is initiated. In machine learning, pipelines are utilized to organize and streamline the workflow of data preprocessing and model training, ensuring a seamless flow of data through each step. The first mentioned component, StandardScaler, is a widely employed preprocessing technique used to standardize the features of a dataset. By scaling the features to have zero mean and unit variance, it ensures similarity in scale, thereby enhancing the performance and stability of certain machine learning models. The second component employed is the regression model, which is used for predicting a continuous target variable based on input features in supervised learning tasks. In this case, the regression model is utilized to predict the price of diamonds using diamond features as inputs. Different regressors, such as linear regression, decision trees, random forests, KNeighbour, and XGB regressor, can be utilized based on the specific problem and dataset characteristics. By combining the StandardScaler and the regression model into a pipeline, the data is transformed using the scaler and subsequently fed into the regression model for training or prediction. This pipeline facilitates the consistent application of the same preprocessing steps and model to new, unseen data in an efficient and convenient manner.

```
# Building pipelins of standard scaler and model for varios regressors.

pipeline_lr=Pipeline([("scalar1",StandardScaler()),
                      ("lr_classifier",LinearRegression())])

pipeline_dt=Pipeline([("scalar2",StandardScaler()),
                      ("dt_classifier",DecisionTreeRegressor())])

pipeline_rf=Pipeline([("scalar3",StandardScaler()),
                      ("rf_classifier",RandomForestRegressor())])


pipeline_kn=Pipeline([("scalar4",StandardScaler()),
                      ("rf_classifier",KNeighborsRegressor())])


pipeline_xgb=Pipeline([("scalar5",StandardScaler()),
                      ("rf_classifier",XGBRegressor())])

# List of all the pipelines
pipelines = [pipeline_lr, pipeline_dt, pipeline_rf, pipeline_kn , pipeline_xgb]

# Dictionary of pipelines and model types for ease of reference
pipe_dict = {0: "LinearRegression", 1: "DecisionTree", 2: "RandomForest", 3: "KNeighbors", 4: "XGBRegressor"}

# Fit the pipelines
for pipe in pipelines:
    pipe.fit(X_train, y_train)
```

Figure 15: Constructing pipeline for efficient Linear Regression.

Step 11:

Finally, results for multiple linear regression are calculated. In Python, **cv_score (cross-validation score)** is commonly used. The model is trained on a subset of the data called the training set and then evaluated on the remaining portion of the data called the validation set. This process is repeated multiple times, with different subsets serving as the validation set each time. The cross-validation score is then calculated as the average performance across all the validation sets. The cross-validation score provides a more robust and reliable estimate of the model's performance compared to evaluating it on a single validation set.

```
cv_results_rms = []
for i, model in enumerate(pipelines):
    cv_score = cross_val_score(model, X_train,y_train,scoring="neg_root_mean_squared_error", cv=10)
    cv_results_rms.append(cv_score)
    print("%s: %f " % (pipe_dict[i], cv_score.mean()))
```

```
LinearRegression: -1348.811824
DecisionTree: -749.347796
RandomForest: -547.658679
KNeighbors: -823.656082
XGBRegressor: -545.458108
```

Figure 16: Cross Validation Score of different regressors.

1. LinearRegression: -1348.811824 The LinearRegression model achieved a mean squared error of approximately 1348.81 during cross-validation. This value represents the average squared difference between the predicted and actual prices of diamonds. Lower values indicate better performance.
2. DecisionTree: -749.347796 The DecisionTree model achieved a mean squared error of approximately 749.35 during cross-validation. This indicates that the DecisionTree model performed better than the LinearRegression model, as it had a lower mean squared error.
3. RandomForest: -547.658679 The RandomForest model achieved a mean squared error of approximately 547.66 during cross-validation. The RandomForest model performed better than both the LinearRegression and DecisionTree models, as it had a lower mean squared error.
4. KNeighbors: -823.656082 The KNeighbors model achieved a mean squared error of approximately 823.66 during cross-validation. This suggests that the KNeighbors model performed worse than the DecisionTree and RandomForest models but better than the LinearRegression model.
5. XGBRegressor: -545.458108 The XGBRegressor model achieved a mean squared error of approximately 545.46 during cross-validation. Similar to the RandomForest model, the XGBRegressor model performed better than the LinearRegression and DecisionTree models, as it had a lower mean squared error.

In summary, based on these scores, the RandomForest and XGBRegressor models appear to have better performance in predicting diamond prices compared to LinearRegression, DecisionTree, and KNeighbors models.

Step 12:

Finally, the model is then used to predict prices based on its training data. The code is shown in the figure below.

```
# Model prediction on test data
pred = pipeline_lr.predict(X_test)
```

```
# Model Evaluation
print("R^2:",metrics.r2_score(y_test, pred))
print("Adjusted R^2:",1 - (1-metrics.r2_score(y_test, pred))*(len(y_test)-1)/(len(y_test)-X_test.shape[1]-1))
print("MAE:",metrics.mean_absolute_error(y_test, pred))
print("MSE:",metrics.mean_squared_error(y_test, pred))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

```
R^2: 0.8890105065854332
Adjusted R^2: 0.888936332274842
MAE: 849.3507396470707
MSE: 1741183.6678057092
RMSE: 1319.5391876733745
```

```
# Model prediction on test data
pred = pipeline_xgb.predict(X_test)
```

```
# Model Evaluation
print("R^2:",metrics.r2_score(y_test, pred))
print("Adjusted R^2:",1 - (1-metrics.r2_score(y_test, pred))*(len(y_test)-1)/(len(y_test)-X_test.shape[1]-1))
print("MAE:",metrics.mean_absolute_error(y_test, pred))
print("MSE:",metrics.mean_squared_error(y_test, pred))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

```
R^2: 0.9810847980166805
Adjusted R^2: 0.9810721569817172
MAE: 278.0934000996743
MSE: 296738.36462685897
RMSE: 544.7369682946614
```

Figure 17: Model Prediction and Evaluation.

When evaluating the results of linear regression regressor it is observed:

1. R-squared (R^2): The R-squared value measures the proportion of variance in the target variable that can be explained by the regression model. An R^2 value of 0.889 indicates that approximately 88.9% of the variance in the target variable is accounted for by the model. Higher R^2 values suggest a better fit of the model to the data.

2. Adjusted R-squared: The adjusted R-squared adjusts the R-squared value by the number of predictors in the model and the sample size. It considers the complexity of the model and penalties for including unnecessary predictors. An adjusted R^2 value of 0.889 indicates that the adjusted R-squared is very close to the R-squared value, suggesting that the model's performance is not significantly affected by the number of predictors.

3. Mean Absolute Error (MAE): The MAE represents the average absolute difference between the predicted and actual values of the target variable. In this case, an MAE of 849.35 indicates that, on average, the model's predictions differ from the actual values by approximately 849.35 units. The MAE measures the magnitude of errors without considering their direction.

4. Mean Squared Error (MSE): The MSE measures the average squared difference between the predicted and actual values. It gives higher weights to larger errors compared to MAE. In this case, the MSE is 1,741,183.67, which indicates that, on

average, the squared difference between the predictions and actual values is approximately 1,741,183.67.

5. Root Mean Squared Error (RMSE): The RMSE is the square root of the MSE and provides a measure of the average magnitude of the prediction errors in the same units as the target variable. An RMSE of 1,319.54 suggests that, on average, the predictions differ from the actual values by approximately 1,319.54 units.

Overall, these results indicate that the regression model has a reasonably good fit to the data, as evidenced by the high R-squared value. The MAE, MSE, and RMSE values provide information about the magnitude of the prediction errors, allowing you to assess the model's accuracy and compare it to other models or benchmarks.

Similarly, when results of XGBoost regressor are evaluated:

1. R-squared (R^2): The R-squared value measures the proportion of variance in the target variable that can be explained by the regression model. An R^2 value of 0.981 suggests that approximately 98.1% of the variance in the target variable is accounted for by the model. Higher R^2 values indicate a better fit of the model to the data, indicating that the model explains a significant portion of the variability in the target variable.

2. Adjusted R-squared: The adjusted R-squared adjusts the R-squared value by the number of predictors in the model and the sample size. It takes into account the complexity of the model and penalties for including unnecessary predictors. An adjusted R^2 value of 0.981, which is very close to the R^2 value, indicates that the model's performance is not significantly affected by the number of predictors.

3. Mean Absolute Error (MAE): The MAE represents the average absolute difference between the predicted and actual values of the target variable. In this case, an MAE of 278.09 suggests that, on average, the model's predictions differ from the actual values by approximately 278.09 units. The MAE measures the magnitude of errors without considering their direction.

4. Mean Squared Error (MSE): The MSE measures the average squared difference between the predicted and actual values. It gives higher weights to larger errors compared to MAE. In this case, the MSE is 296,738.36, indicating that, on average, the squared difference between the predictions and actual values is approximately 296,738.36.

5. Root Mean Squared Error (RMSE): The RMSE is the square root of the MSE and provides a measure of the average magnitude of the prediction errors in the same units as the target variable. An RMSE of 544.74 suggests that, on average, the predictions differ from the actual values by approximately 544.74 units.

Overall, these results indicate that the regression model has a very good fit to the data, as evidenced by the high R-squared value of 0.981. The low MAE, MSE, and RMSE values suggest that the model's predictions are close to the actual values, with relatively small errors. This indicates that the model is performing well and providing accurate predictions.

Step 13:

The step of normalizing the data refers to the process of transforming the numerical variables in a dataset to a common scale. This is done to eliminate differences in scales among the variables, ensuring that they are on a similar numerical range.

Normalization helps to prevent variables with larger magnitudes from dominating the analysis and model training process. It is particularly useful when using certain machine learning algorithms that are sensitive to the scale of the input features.

```
In [218…  #Normalizing the Data
          df_norm = df/np.sqrt(np.power(df,2).sum(axis=0))
          df_norm.head()
```

Out[218]:

	carat	cut	color	clarity	depth	table	x	y	z
0	0.001067	0.005302	0.005470	0.001969	0.004287	0.004118	0.002912	0.002931	0.002900
1	0.000974	0.004241	0.005470	0.002954	0.004169	0.004568	0.002868	0.002828	0.002756
2	0.001067	0.002121	0.005470	0.004923	0.003966	0.004867	0.002986	0.002997	0.002756
3	0.001345	0.004241	0.001823	0.003938	0.004350	0.004343	0.003097	0.003115	0.003138
4	0.001438	0.002121	0.000912	0.001969	0.004413	0.004343	0.003200	0.003203	0.003281

Figure 18: Normalizing the data.

Step 14:

The Analytic Hierarchy Process (AHP) is a widely used method for calculating weights or relative priorities in multi-criteria decision-making problems. The process involves a series of pairwise comparisons to determine the importance or priority of criteria and alternatives. For the sake of simplicity, this paper utilizes correlation between price and all other attributes mentioned in the solution to determine priority of attributes. Furthermore, an AHP calculator was utilized, provided by HP-OS author: Klaus D. Goepel, BPMSG. The results and corresponding weights of the solutions are shown below.

Priorities

These are the resulting weights for the criteria based on your pairwise comparisons:

	Cat	Priority	Rank	(+)	(-)
1	Carat	17.2%	4	8.5%	8.5%
2	Cut	3.0%	6	1.3%	1.3%
3	Clarity	1.5%	9	0.7%	0.7%
4	Color	2.2%	8	0.6%	0.6%
5	Depth	2.7%	7	1.1%	1.1%
6	Table	5.5%	5	4.0%	4.0%
7	x	22.6%	1	6.0%	6.0%
8	y	22.6%	1	6.0%	6.0%
9	z	22.6%	1	6.0%	6.0%

Number of comparisons = 36
Consistency Ratio CR = 6.1%

Decision Matrix

The resulting weights are based on the principal eigenvector of the decision matrix:

	1	2	3	4	5	6	7	8	9
1	1	9.00	9.00	7.00	8.00	7.00	0.50	0.50	0.50
2	0.11	1	4.00	1.00	1.00	1.00	0.11	0.11	0.11
3	0.11	0.25	1	0.50	0.25	0.14	0.11	0.11	0.11
4	0.14	1.00	2.00	1	1.00	0.14	0.11	0.11	0.11
5	0.12	1.00	4.00	1.00	1	0.50	0.11	0.11	0.11
6	0.14	1.00	7.00	7.00	2.00	1	0.17	0.17	0.17
7	2.00	9.00	9.00	9.00	9.00	6.00	1	1.00	1.00
8	2.00	9.00	9.00	9.00	9.00	6.00	1.00	1	1.00
9	2.00	9.00	9.00	9.00	9.00	6.00	1.00	1.00	1

Principal eigen value = 9.703
Eigenvector solution: 6 iterations, delta = 1.9E-8

Figure 19: Priorities of features based on AHP method.

```
In [17]: #Defining weights based on AHP method
         W = 0.18, 0.04, 0.015, 0.02, 0.02, 0.06, 0.22, 0.22, 0.225
```

Figure 20: Defining Weights based on AHP Analysis.

Step 15:

The step of formulating the weighted normalized matrix involves integrating the normalized data with the calculated weights for each criterion. In this context, the weights are derived using an Analytical Hierarchy Process (AHP) calculator, which helps determine the relative importance of different criteria. Additionally, the correlation matrix obtained from the regression analysis is utilized to assign priority to each feature.

The process can be summarized as follows:

1. Normalization: The numerical attributes of the dataset are normalized using appropriate techniques to ensure consistent scales and eliminate differences in magnitudes.
2. Weight Calculation: The AHP calculator is employed to determine the relative weights for each criterion. This process involves evaluating pairwise comparisons and synthesizing the judgments to derive the importance or priority of each criterion.
3. Correlation Matrix: The correlation matrix obtained from the regression analysis is utilized to assess the relationships between features. This matrix provides valuable insights into the strength and direction of the linear associations among variables.
4. Assigning Priority: Based on the correlation matrix, priority is assigned to each feature. Features that exhibit stronger correlations with the target variable or demonstrate higher predictive power are given higher priority in the analysis.
5. Weighted Normalized Matrix: The normalized data is multiplied element-wise by their corresponding weights to incorporate the relative importance of each criterion. This results in a weighted normalized matrix that reflects the combined influence of both the normalization process and the assigned weights.

By formulating the weighted normalized matrix, the analysis incorporates both the normalized data and the prioritized criteria, enabling a comprehensive assessment of the data based on their relative importance and interrelationships.

```
#Defining and Calculating weighted normal matrix
df_norm_w = df_norm * w
df_norm_w.head()
```

	carat	cut	color	clarity	depth	table	x	y	z
0	0.000192	0.000212	0.000082	0.000039	0.000086	0.000247	0.000641	0.000645	0.000652
1	0.000175	0.000170	0.000082	0.000059	0.000083	0.000274	0.000631	0.000622	0.000620
2	0.000192	0.000085	0.000082	0.000098	0.000079	0.000292	0.000657	0.000659	0.000620
3	0.000242	0.000170	0.000027	0.000079	0.000087	0.000261	0.000681	0.000685	0.000706
4	0.000259	0.000085	0.000014	0.000039	0.000088	0.000261	0.000704	0.000705	0.000738

Figure 21: Weighted Normalized Matrix.

Step 16:

The step of identifying the positive and negative ideals for TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) analysis involves determining the reference points that represent the best and worst possible values for each criterion in the decision matrix. The distances calculated using the positive and negative ideal solutions help quantify the similarity or proximity of each diamond to the ideal solutions and facilitate the ranking and evaluation process in TOPSIS analysis.

```
#Identifying positive and negative ideals for TOPSIS
positive_ideal = df_norm_w.max()
negative_ideal = df_norm_w.min()
print(positive_ideal, negative_ideal)
```

```
carat      0.004184
cut        0.000212
color      0.000096
clarity    0.000158
depth      0.000110
table      0.000427
x          0.001742
y          0.009542
z          0.008538
dtype: float64 carat      0.000167
cut        0.000042
color      0.000014
clarity    0.000020
depth      0.000060
table      0.000193
x          0.000000
y          0.000000
z          0.000000
dtype: float64
```

Figure 22: Positive Ideal and Negative Ideal solutions for each attribute.

Step 17:

The step of defining separation measurements for the positive and negative ideal solution involves calculating the distances or separations between each alternative (diamond specimen) and the positive and negative ideal solutions. These separation measurements quantify the similarity or proximity of each alternative to the respective ideal solution.

1. Positive Ideal Solution: The positive ideal solution represents the best possible values for each criterion in the decision matrix. To calculate the separation measurements for each alternative with respect to the positive ideal solution, the Euclidean distance or any other distance metric can be used. The Euclidean distance between an alternative and the positive ideal solution measures the straight-line distance between them in the multi-dimensional space of criterion values. The smaller the distance, the closer the alternative is to the positive ideal solution, indicating a higher performance or desirability.
2. Negative Ideal Solution: The negative ideal solution represents the worst possible values for each criterion in the decision matrix. Similar to the positive ideal solution, the separation measurements for each alternative with respect to the negative ideal solution can be computed using a distance metric, such as the Euclidean distance. Again, a smaller distance indicates that the alternative is closer to the negative ideal solution, suggesting a lower performance or desirability.

```
#Doing seperation measurements
#Positive idea
SM_P = np.sqrt(np.power(df_norm_w - positive_ideal, 2).sum(axis=1))
#Negative idea
SM_N = np.sqrt(np.power(df_norm_w - negative_ideal, 2).sum(axis=1))
```

```
print(SM_N)
```

```
0        0.001136
1        0.001095
2        0.001129
3        0.001210
4        0.001246
           ...
53935    0.001689
53936    0.001692
53937    0.001673
53938    0.001820
53939    0.001733
Length: 53940, dtype: float64
```

Figure 23: Separation Measurement for each specimen based on Ideal Solutions.

By calculating the separation measurements for both the positive and negative ideal solutions, the TOPSIS analysis quantifies the relative proximity or similarity of each alternative to these reference points. These separation measurements serve as a basis for ranking the alternatives and determining their overall performance or suitability in relation to the ideal solutions.

Step 18:

The step of ranking each specimen based on their separation measurements and sorting them according to their ranking is a crucial part of the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) analysis. This process allows for the identification of the most suitable or preferable specimens based on their proximity to the ideal solutions.

```
#Ranking each specimen based on the criteria and weights
final_rank = SM_N /(SM_N + SM_P)
```

```
print(final_rank)
```

```
0        0.082740
1        0.079775
2        0.082233
3        0.088171
4        0.090870
           ...
53935    0.122893
53936    0.123177
53937    0.121816
53938    0.132306
53939    0.126092
Length: 53940, dtype: float64
```

```
#Sorting all the specimens based on their ranking
final_rank.sort_values(ascending=False)
```

```
24067    0.592503
48410    0.475036
49189    0.368909
27415    0.329654
27630    0.309142
           ...
26243    0.060715
15951    0.056486
11963    0.048372
49557    0.031733
49556    0.031733
Length: 53940, dtype: float64
```

Figure 24: Calculated Rank of each specimen based on TOPSIS analysis and sorting them.

By ranking and sorting the specimens, the TOPSIS analysis provides a clear order of preference or suitability based on their proximity to the ideal solutions. The ranking and sorting step enables decision-makers to identify and focus on the top-ranked specimens that exhibit the highest level of similarity to the positive ideal solution and the lowest level of similarity to the negative ideal solution.

9 Conclusion:

This paper sets out with the objective of formulating a method to rate specimens of diamond based on measurable attributes. In many business and even some scientific environments, the use of data to consolidate prior knowledge into a useable tool has become a necessity. Jewel owners, collectors and jewelry shop owners have historically relied on experiential knowledge of an elder for the ultimate judgement of a specimen of a diamond. In the modern era, this reliance on one person's knowledge and judgement is not sufficient for a business to remain competitive globally. This calls for a system to accurate price diamonds which relies on measurable attributes and can be transformed easily into a price valuation of any specimen. For this exact purpose, this paper utilizes the historical data for pricing of a diamond specimen to allocate weights and priorities to attributes like the size, cut, color and clarity of a diamond. These weights are then used in the ranking of each specimen in order to determine its price.

This paper utilized regression analysis and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method to estimate the price of diamond specimens. The regression analysis helped identify the relationships between the diamond attributes and their prices, providing valuable insights into the factors influencing diamond pricing.

The correlation between features of a diamond specimen and its price provides a statistical viewpoint on what affects the price of a diamond. These correlations can then be used to prioritize features to determine their weight according to Analytical Hierarchy Process (AHP).

The TOPSIS analysis further enhanced the estimation process by considering multiple criteria and determining the proximity of each diamond specimen to the positive and negative ideal solutions. This allowed for the ranking and sorting of the specimens based on their performance relative to the ideal solutions, providing a systematic approach to identifying the most suitable diamond specimens.

Through the application of these methods, this research successfully developed a comprehensive framework for estimating diamond prices. By considering both the regression analysis and the TOPSIS analysis, a more robust and accurate estimation of diamond prices was achieved.

The calculations performed not only give an idea about which attribute affects the price of a diamond the most, but also how a combination of different attributes translates to the overall valuation of a diamond. As observed in *figure 20*, the carat of a diamond seems to affect the price the most, followed by the x, y and z dimensions. In order to get these rankings, the AHP model compares each attribute against the price in order to determine its relative importance as depicted in figure 18, the pairwise comparison matrix.

Finally, each specimen of diamond in our database is then ranked according to TOPSIS analysis, where the ideal solution of each attribute is calculated and then every single specimen's separation distance from the positive and negative ideal solution is calculated as shown in *figure 24*. These ratings are then used to rank each specimen and determine its value. The highest ranked specimen of diamond is the priciest one. This formulation can now be used to determine the ranking of new specimens and place them on the ranking chart in order to evaluate their price.

It is important to acknowledge the limitations of the study. The accuracy of the price estimates depends on the quality and representativeness of the dataset used, as well as the assumptions made during the regression analysis and TOPSIS analysis. Additionally, the interpretation and generalizability of the results should be done cautiously, as the study focused on a specific dataset and may not be applicable to all diamond specimens.

Future research can focus on expanding the dataset, incorporating additional criteria, or exploring alternative machine learning algorithms for regression analysis. Furthermore, the inclusion of expert opinions or customer preferences can further enhance the estimation process.

Overall, this study contributes to the field of diamond pricing estimation by combining regression analysis and TOPSIS analysis. The framework developed can assist diamond industry professionals, researchers, and consumers in making informed decisions based on a more comprehensive understanding of diamond pricing factors.

10 References

Bhutia, P. W., & Phipon, R. (2012). Appication of ahp and topsis method for supplier selection problem. *IOSR Journal of Engineering*, *2*(10), 43–50. www.iosrjen.org43%7CPage

Bijaya, L., Pradhan, A., Tu, N., Subedi, G., Kapil, M., Subedi, D., & Prof, A. (2019). *Correlation and Regression Analysis Using SPSS*. http://www.oxfordcollege.edu.np

Diamond Characteristics: The Four Cs. (n.d.). Retrieved June 20, 2023, from https://www.germanjoyero.com/en/diamond-characteristics/

Diamond Prices: How to Calculate a Diamond's Value & Worth. (n.d.). Retrieved June 20, 2023, from https://www.diamonds.pro/education/diamond-prices/

Diamond Quality Factors. (n.d.). Retrieved June 20, 2023, from https://www.gia.edu/diamond-quality-factor

Diamonds | Kaggle. (n.d.). Retrieved June 20, 2023, from https://www.kaggle.com/datasets/shivam2503/diamonds

Douglas C. Montgomery, Elizabeth A. Peck, G. G. V. (n.d.). *Introduction to Linear Regression Analysis - Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining - Google Books*. Retrieved June 20, 2023, from https://books.google.com.tr/books?hl=en&lr=&id=tCIgEAAAQBAJ&oi=fnd&pg=PR13&dq=regression+analysis&ots=lfzeUwlXMq&sig=65eWDFn9RJ7SAOMW1nmW4C1TXVU&redir_esc=y#v=onepage&q=regression analysis&f=false

Hansen, P., & Devlin, N. (2019). Multi-Criteria Decision Analysis (MCDA) in Healthcare Decision-Making. *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/ACREFORE/9780190625979.013.98

Jahan, A., & Edwards, K. L. (2013). Multi-criteria Decision-Making for Materials Selection. *Multi-Criteria Decision Analysis for Supporting the Selection of Engineering Materials in Product Design*, 31–41. https://doi.org/10.1016/B978-0-08-099386-7.00003-9

Kildiene, E. K. Z. Z. T. S. (2014). *State of the art surveys on MCDM methods - Edmundas Kazimieras ZAVADSKAS, Zenonas TURSKIS, Simona KILDIENĖ.pdf*.

Köksalan, M., Wallenius, J., & Zionts, S. (2011). Multiple Criteria Decision Making:From Early History to the 21st Century. *World Scientific Books*, 1–198. https://doi.org/10.1142/8042

Kumar, A., Sah, B., Singh, A. R., Deng, Y., He, X., Kumar, P., & Bansal, R. C. (2017). A review of multi criteria decision making (MCDM) towards sustainable renewable energy development. *Renewable and Sustainable Energy Reviews*, *69*, 596–609. https://doi.org/10.1016/J.RSER.2016.11.191

Liang, K. Y., & Zeger, S. L. (2003). Regression Analysis for Correlated Data. *Https://Doi.Org/10.1146/Annurev.Pu.14.050193.000355*, *14*, 43–68. https://doi.org/10.1146/ANNUREV.PU.14.050193.000355

Marzouk, M., & Sabbah, M. (2021). AHP-TOPSIS social sustainability approach for selecting supplier in construction supply chain. *Cleaner Environmental Systems*, *2*.

https://doi.org/10.1016/J.CESYS.2021.100034

Multi-Criteria Decision Analysis (MCDA/MCDM) | 1000minds. (n.d.). Retrieved January 16, 2023, from https://www.1000minds.com/decision-making/what-is-mcdm-mcda

Multiple Criteria Decision Making | International Society on MCDM. (n.d.). Retrieved January 16, 2023, from http://www.mcdmsociety.org/

Norman R. Draper, H. S. (n.d.). *Applied Regression Analysis - Norman R. Draper, Harry Smith - Google Books*. Retrieved June 20, 2023, from https://books.google.com.tr/books?hl=en&lr=&id=d6NsDwAAQBAJ&oi=fnd&pg=PR13&dq=regression+analysis&ots=Byuak9nXOQ&sig=ghp_5BLH8CR7P38haN2a34KAvp0&redir_esc=y#v=onepage&q=regression analysis&f=false

Pal, O., Gupta, A. K., & Garg, R. K. (2013). Supplier Selection Criteria and Methods in Supply Chains: A Review. *International Journal of Economics and Management Engineering*, *7*(10), 2667–2673. https://doi.org/10.5281/ZENODO.1088140

Podvezko, V. (2009). Application of AHP technique. *Journal of Business Economics and Management - J BUS ECON MANAG*, *10*, 181–189. https://doi.org/10.3846/1611-1699.2009.10.181-189

Regression Analysis - Rudolf J. Freund, William J. Wilson, Ping Sa - Google Books. (n.d.). Retrieved June 20, 2023, from https://books.google.com.tr/books?id=Us4YE8lJVYMC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26. https://doi.org/10.1016/0377-2217(90)90057-I

Samprit Chatterjee, A. S. H. (n.d.). *Regression Analysis by Example - Samprit Chatterjee, Ali S. Hadi - Google Books*. Retrieved June 20, 2023, from https://books.google.com.tr/books?hl=en&lr=&id=uiu5XsAA9kYC&oi=fnd&pg=PR1&dq=regression+analysis&ots=mqIxOac9Om&sig=FyFsEwVNCbFpab7NCCGBN01vIKc&redir_esc=y#v=onepage&q=regression analysis&f=false

Sarkar, B. (2010). Fuzzy decision making and its applications in cotton fibre grading. *Soft Computing in Textile Engineering*, 353–383. https://doi.org/10.1533/9780857090812.5.353

Sykes, A. O. (n.d.). *An Introduction to Regression Analysis*. Retrieved June 20, 2023, from https://chicagounbound.uchicago.edu/law_and_economics

Zavadskas, E. K., Mardani, A., Turskis, Z., Jusoh, A., & Nor, K. M. (2016). Development of TOPSIS Method to Solve Complicated Decision-Making Problems — An Overview on Developments from 2000 to 2015. *Https://Doi.Org/10.1142/S0219622016300019*, *15*(3), 645–682. https://doi.org/10.1142/S0219622016300019