A. Gary Anderson School of Management

School of Business

MGT 286A : Capstone in Business Analytics

Final Project Report

Team:

Shail Desai (862071746)

Richard AllenPaul (862467207)

Sahana RajeshLal (862465869)

Introduction to Cubic Zirconia

Cubic Zirconia (CZ) is a synthetic gemstone that closely resembles diamonds. Due to its affordability, durability, and visual similarity to diamonds, CZ has become a popular choice for consumers seeking the aesthetic appeal of diamonds without the associated cost. This growing market demand positions Cubic Zirconia as a key product for manufacturers in the gemstone industry.

Strategic Importance of Price Prediction

For Gem Stones Co. Ltd, a company specializing in the manufacture of Cubic Zirconia, the ability to accurately predict the price of each stone based on its attributes is critical for several reasons:

1. Optimization of Profit Margins: By understanding which attributes of Cubic Zirconia contribute to higher prices, the company can focus its manufacturing process on producing stones that are more likely to yield higher profits. This approach allows for more efficient resource allocation and inventory management.

2. Market Competitiveness: In a competitive market, pricing strategies that are informed by detailed analysis can provide a significant advantage. Price prediction enables the company to set competitive prices that appeal to consumers while ensuring profitability.

3. Targeted Marketing Strategies: Knowing the price and value of different stones allows for more targeted marketing campaigns. The company can segment its market more effectively, targeting specific customer groups with products that meet their price and quality expectations.

4. Enhanced Customer Satisfaction: By aligning production with consumer demand for certain qualities in Cubic Zirconia, the company can enhance customer satisfaction. This alignment helps in building brand loyalty and reputation in the market.

5. Investment and Expansion Decisions: Accurate price prediction aids in making informed decisions regarding investments in new technologies and expansion into new markets. It helps in assessing the potential return on investment for different strategic initiatives.

Problem Statement

The analysis to predict the price of Cubic Zirconia based on various attributes is not just a technical exercise but a strategic necessity for Gem Stones Co. Ltd. It supports the company in navigating the complexities of the market, optimizing operations, and enhancing profitability. Identifying the top attributes that influence price will empower the company to make informed decisions across various aspects of its business, from production to marketing, thus securing a competitive edge in the gemstone industry.

Data Dictionary

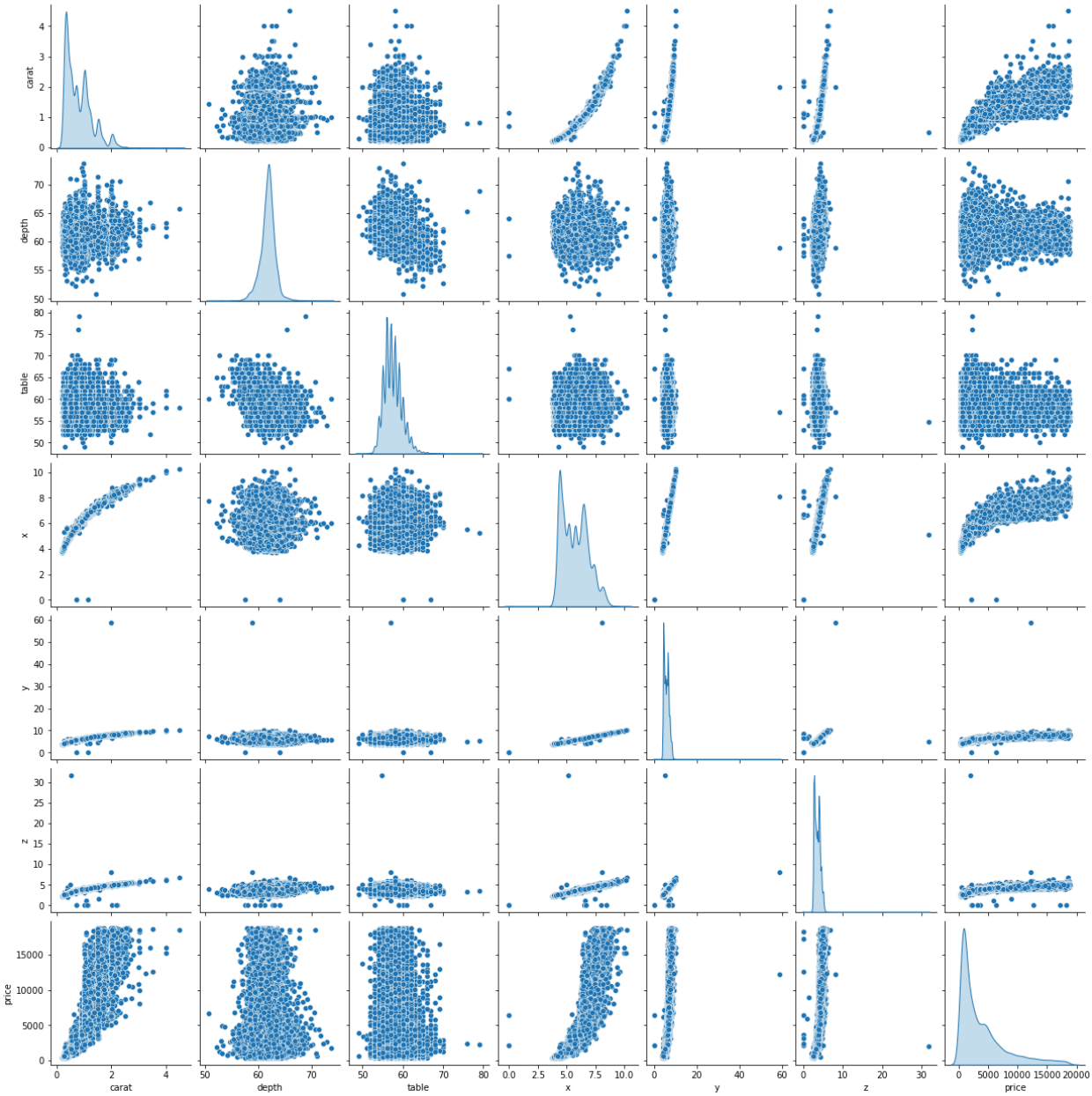
A comprehensive data dictionary details the variables considered in the analysis:

Carat	Weight of the cubic zirconia.
Cut	Quality of the cubic zirconia's cut, ranked from Fair to Ideal.
Color	Color grading of the cubic zirconia, from D (best) to J (worst).
Clarity	Refers to the presence of inclusions and blemishes, ranked from Flawless to I3.
Depth	Height of the cubic zirconia measured from the culet to the table, divided by its average girdle diameter.
Price	The price of the cubic zirconia
X,Y,Z	Dimensions of the cubic zirconia in millimeters (length, width, height).

Summary of the Dataset

- The dataset had 26967 rows and 10 columns
- The Unnamed column which replicated the index column was dropped
- The dataset was checked for null values and it was found that the depth column had 697 null values
- There are no duplicated rows in the dataset
- The unique number of values in each column is as follows

Pairplot showing the Distribution of variables



Exploratory Data Analysis

1. **Collective histogram :**

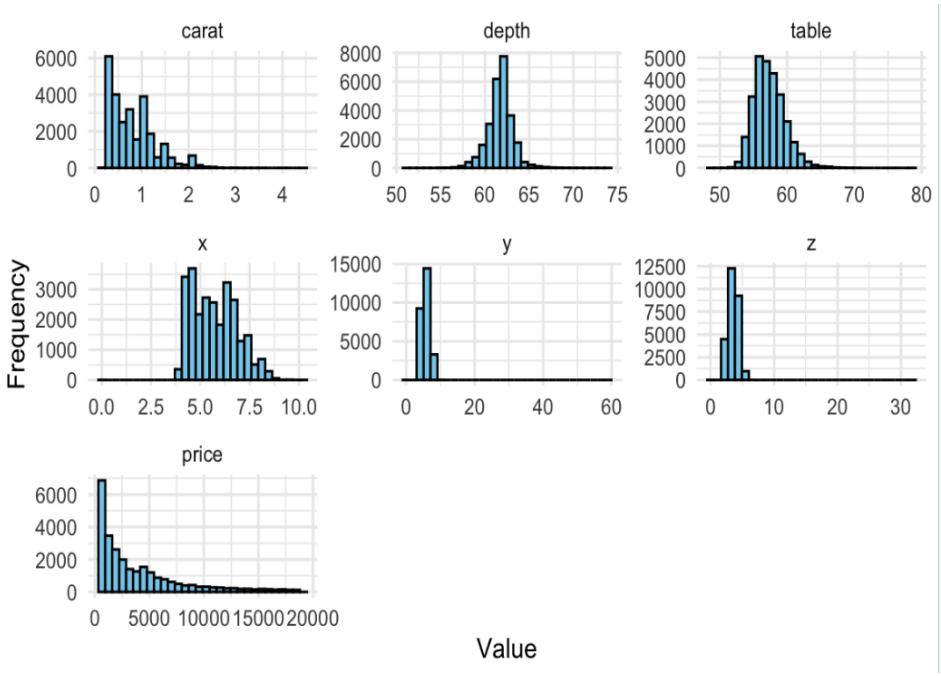
- The collective histogram reveals the distribution of carat, weight, depth, and price of diamond characteristics in a dataset.

Carat : This histogram shows the distribution of the diamond weight measured in carats. Most diamonds in this dataset seem to be less than 1 carat, with frequencies decreasing as the carat size increases.

Depth : The distribution here is normal depicting that most diamonds have a similar proportionate depth.

Table : This refers to the width of the diamond's table expressed as a percentage of its average diameter. Again, we see a normal distribution around a central value, indicating a commonality in table size proportion across these diamonds.

X,Y,Z : These histograms represent the length(x), width(y) and depth(z) in millimeters of each diamond. The 'x' and 'y' distributions are fairly similar and skewed to the right, which means most of the diamonds have smaller lengths and widths, whereas 'z' shows a sharp peak, suggesting a standardization in diamond cut depth.
- Price : This shows the price of the diamonds, which is heavily right-skewed, meaning a large number of diamonds are on the lower end of the price spectrum, with fewer diamonds at higher prices.



2. Analysis of color, cut, and clarity against frequency :

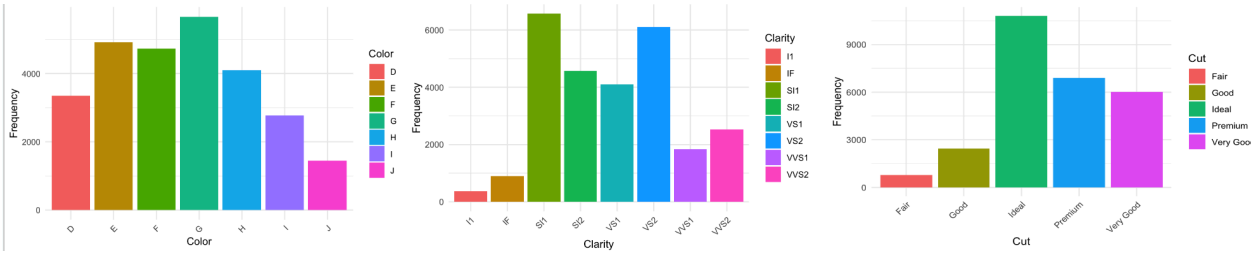
Analysis of color, cut, and clarity against frequency demonstrates consumer preferences for near-colorless stones (G and H color grades), slight inclusions (SI1 clarity), and optimal brilliance cut (ideal).

Color: Most diamonds are graded G and H, showing a preference for near colorless stones. D and J colors are the least common, with D, E, F being colorless and less frequent.

Clarity: SI1 is the top clarity grade chosen, indicating a lean towards diamonds with minor inclusions. The highest (IF) and lowest (I1) clarity grades are rare, suggesting people opt for diamonds that balance imperfections with cost.

Cut: 'Ideal' cut diamonds are the most popular, reflecting a preference for the best brilliance and proportion. The 'Fair' cut is the least favored, highlighting that quality in cut is a priority.

Overall Preference: Consumers prefer near colorless diamonds, are somewhat flexible with clarity, often choosing diamonds with slight inclusions, and prioritize high-quality cuts for optimal brilliance.



3. Price analysis :

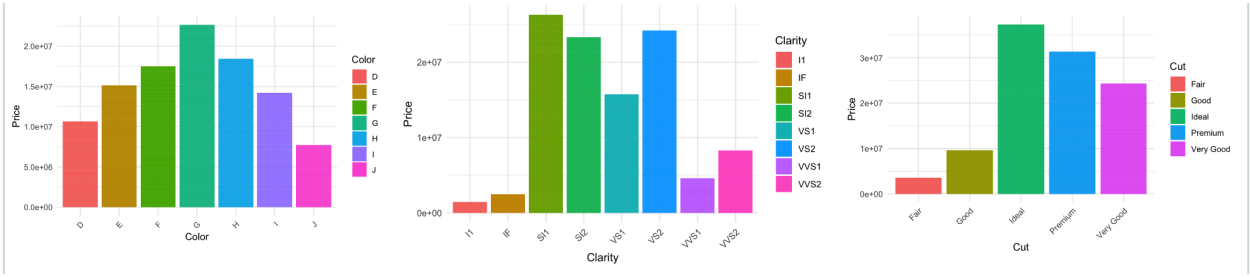
Price analysis suggests near colorless diamonds (G and H) are priced highest due to abundance over rarer colorless (D-F) diamonds, and 'Ideal' cut diamonds command the highest prices for their exceptional brilliance.

Color Impact: Near colorless diamonds (G and H) have the highest total price, reflecting their abundance over the rarer colorless diamonds (D-F).

Clarity Trends: SI1 clarity diamonds, with slight inclusions, top the total price chart, suggesting a market favor for their balance of quality and availability.

Cut Preferences: 'Ideal' cut diamonds boast the highest total price, showing a consumer preference for their exceptional brilliance and light performance.

Market Insights: The highest total prices are commanded by the most abundant qualities, indicating a market trend towards diamonds that balance quality with availability.



4. Boxplots :

A boxplot analysis indicates a wide range of carat sizes, mostly under 2 carats, with a median price around $5000, highlighting significant price variation influenced by quality measures.

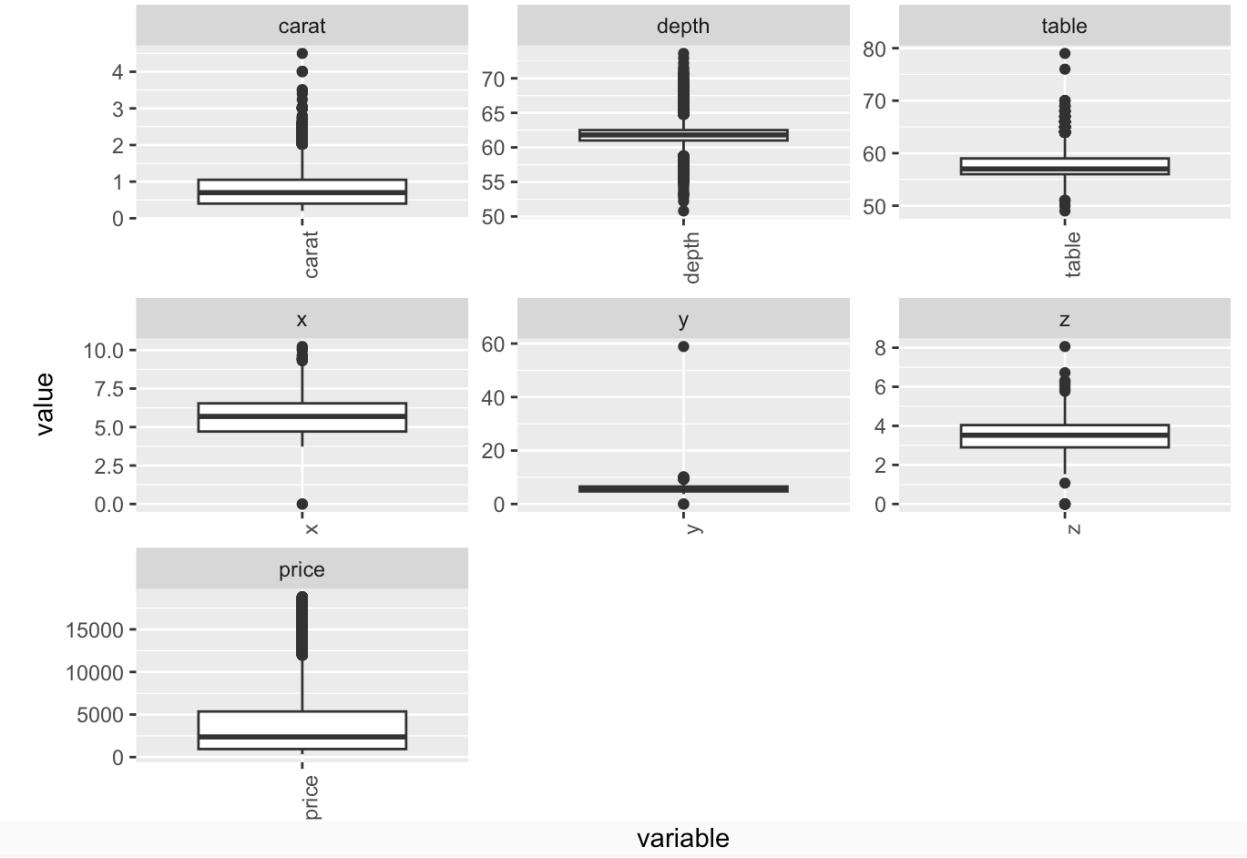
Carat: Shows a median close to 1 with many outliers, suggesting a wide range of gem sizes, mostly smaller than 2 carats.

Depth: Concentrated around 60 to 65, with a few outliers, indicating that most gems have a standard depth percentage.

Table: Distribution is tight around 55 to 60 with some extreme outliers, indicating a consistent table size with few exceptions.

Dimensions (x, y, z): The spread is similar across these dimensions, indicating a consistent proportionality, but 'y' shows an extreme outlier, which may need investigation.

Price: Varies widely, with a median around $5,000 and many high-value outliers, reflecting a significant variation in gem prices, potentially influenced by their carat and other quality measures.

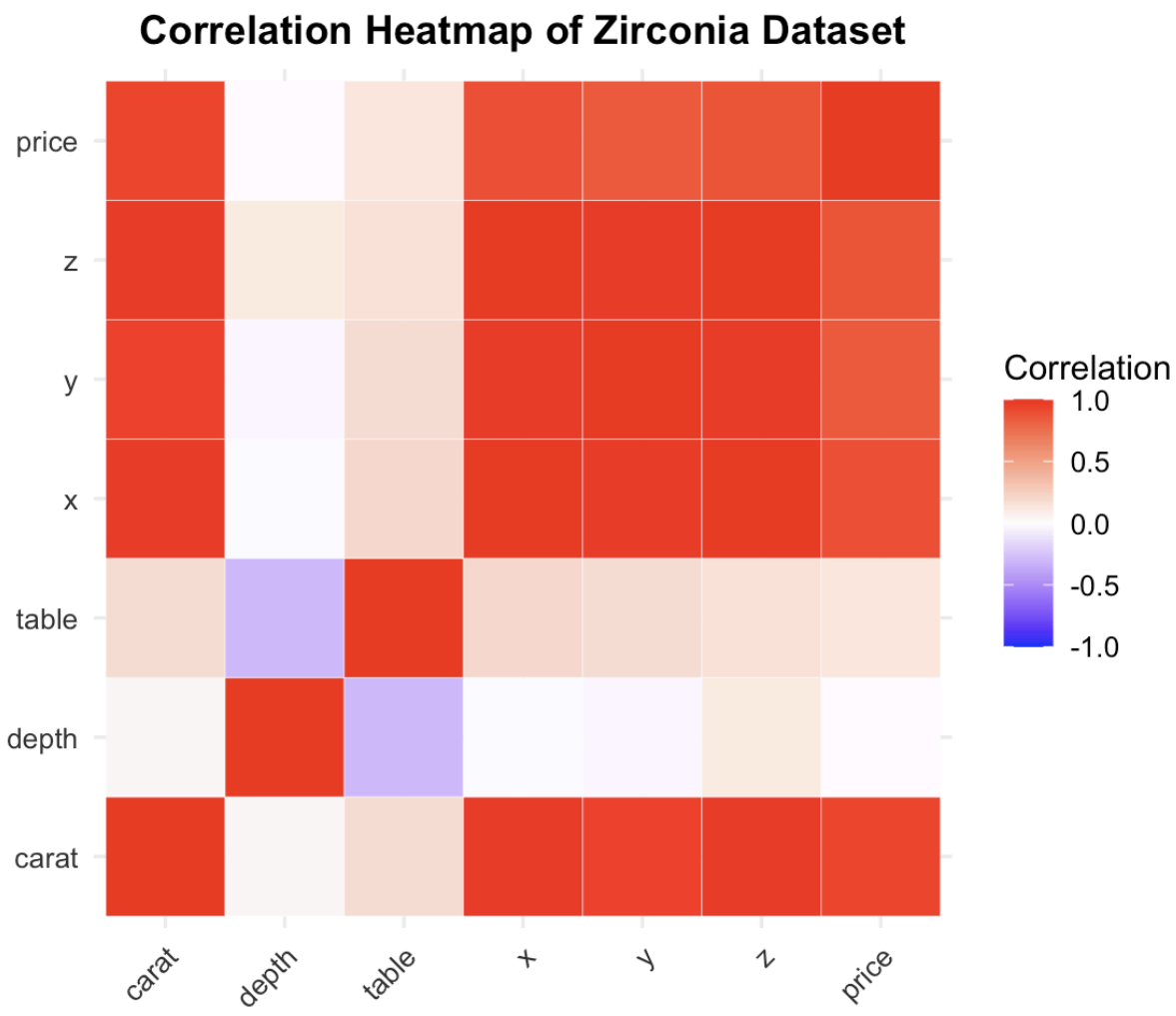


5. Correlation Matrix & Bivariate Analysis :

The correlation matrix reveals strong positive correlations between dimensions (X, Y, Z) with each other and with carat and price, indicating larger stones are generally more expensive. Table size shows a slight negative correlation with depth, suggesting variations in stone cut design. Each square shows the correlation coefficient between variables on the x-axis and y-axis. The coefficient ranges from-1 to 1 where:

- 1 indicates a perfect positive correlation(as one variable increases, the other does too), indicates no correlation,
- 1 indicates a perfect negative correlation(as one variable increases, the other decreases. The intensity of the color represents the strength of the correlation, with red indicating positive, blue indicating negative, and white representing no correlation. For example, the squares along the diagonal are red because they represent the correlation of each variable with itself, which is always perfect. It looks like ‘carat’ is positively correlated with ‘x’,’y’,’z’ and ‘price’, suggesting that as the carat size increases, the physical dimensions and price of the zirconia also increases. Conversely, there might be a negative correlation indicated by blue between ‘table’ and some other variables, but the details are not clear. Ideal

cut seems to be the most preferred because of its lower prices.G is the most preferred color dues its affordable price whereas J is the least due to its higher prices.



Data Preprocessing

Test-Train Split

For Gem Stones Co. Ltd, the dataset containing nearly 27,000 records of Cubic Zirconia attributes and their prices undergoes a fundamental preprocessing step known as the test-train split. This technique divides the dataset into two parts:

Training Set (80%): This larger portion of the data is used to train the machine learning model. It includes a variety of scenarios and attributes that allow the model to learn the underlying patterns and relationships between the attributes of Cubic Zirconia stones and their prices.

Test Set (20%): This smaller portion is used to evaluate the model's performance. It is kept separate from the training process to provide an unbiased assessment of how well the model can generalize to new, unseen data.

The decision to use an 80/20 split for dividing the dataset into training and testing sets is based on several considerations:

- Balance Between Learning and Validation: This split provides a substantial amount of data for the model to learn from, while still reserving enough instances for a robust validation of the model's performance.
- Reduction of Overfitting: Training on 80% of the data helps in reducing the risk of overfitting, where the model performs well on the training data but poorly on new data.
- Efficient Use of Data: Given the substantial size of the dataset, the 80/20 split ensures that the model has enough examples to learn from, without significantly compromising the ability to test the model's generalization.

Model Building: Approaching the Problem Statement

In addressing the problem statement set forth by Gem Stones Co. Ltd, three distinct models were employed: linear regression, decision tree regression, and random forest regression. Each of these models contributes uniquely towards predicting the price of cubic zirconia stones, thereby aiding in distinguishing between higher and lower profitable stones for better profit share optimization.

Linear Regression: This model establishes a linear relationship between the dependent variable (price) and one or more independent variables (attributes of cubic zirconia). It is particularly effective in identifying the direct impact of each attribute on the stone's price. By quantifying these relationships, linear regression aids in understanding how individual features contribute to the price, providing a clear basis for price prediction and profit analysis.

Code

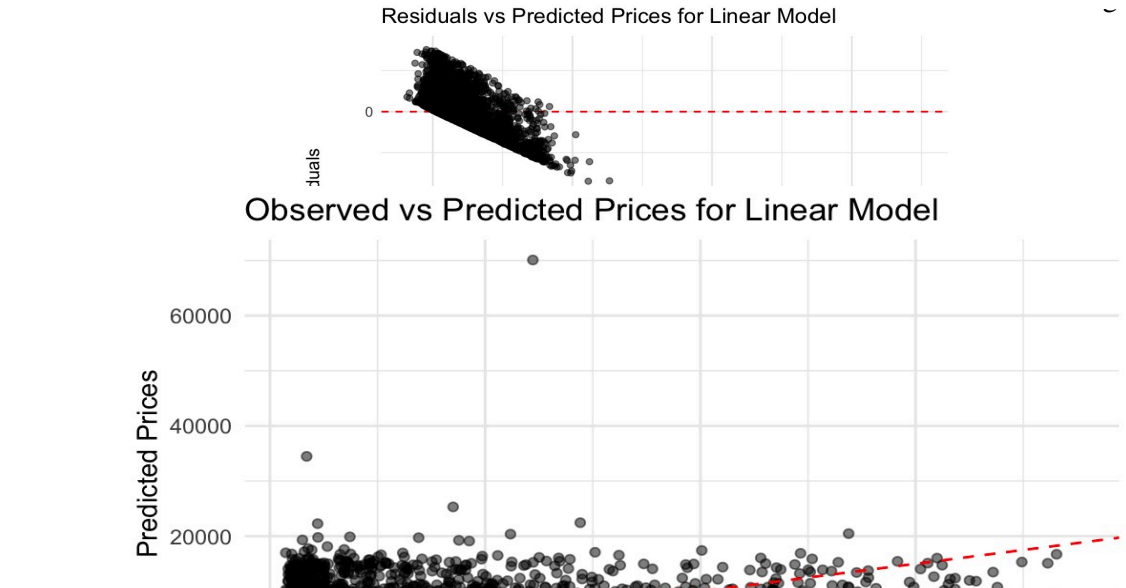
```
# Fitting Linear Regression Model
lm_model <- lm(price ~ ., data = train_data)

# Linear Regression Summary
lm_summary <- summary(lm_model)
print(lm_summary)

# After printing the summary, you can also extract the coefficients
lm_coefficients <- lm_summary$coefficients
print(lm_coefficients)
```

Output

```
Residual standard error: 1116 on 21004 degrees of freedom
  (546 observations deleted due to missingness)
Multiple R-squared:  0.923,     Adjusted R-squared:  0.9229
F-statistic: 1.048e+04 on 24 and 21004 DF,  p-value: < 2.2e-16
```



Model Findings and Business Implications

Residuals Analysis: The first plot shows the residuals (differences between observed and predicted prices) against the predicted prices. The random scatter of points suggests that the model's errors are distributed without a clear pattern, which is a good indication of model appropriateness. However, the presence of a few points with high residuals indicates potential outliers or extreme values that the model does not predict well. From a business perspective, these outliers might represent unique stones with attributes not well represented in the dataset or special market conditions, prompting a review of product quality or pricing strategy for these exceptions.

Predictive Accuracy: The second plot illustrates the relationship between observed and predicted prices. The red dashed line indicates the ideal fit, and the points represent actual predictions. The close clustering of points along the line suggests high predictive accuracy for most stones. In business terms, this indicates the company can trust the model to set prices that closely mirror market expectations, reducing the risk of underpricing or overpricing their products.

Model Performance Metrics: The summary statistics show a high Multiple R-squared (0.923) and Adjusted R-squared (0.9229), indicating that the model explains over 92% of the variance in cubic zirconia prices. This high level of explanatory power is a strong indication that the model's predictions are reliable. The very low p-value (< 2.2e-16) suggests that the model's parameters

are highly significant. For Gem Stones Co. Ltd, this means the model is a powerful tool for making informed pricing decisions, as it has statistically significant predictive power.

Residual Standard Error: The residual standard error (1116 on 21004 degrees of freedom) provides a measure of the typical size of the prediction errors. While it is relatively low, any effort to reduce this error can further enhance pricing accuracy and profitability.

The effectiveness of the linear model in predicting prices allows Gem Stones Co. Ltd to strategically adjust operations to prioritize the production of attributes that lead to higher prices and, consequently, higher profit margins.

Model 2 - Random forest classifier

As an ensemble method that utilizes multiple decision trees to improve prediction accuracy, random forest regression reduces the risk of overfitting—a common challenge in model building. By aggregating the predictions from multiple decision trees, it provides a more robust and generalized model. This method enhances the accuracy of price predictions by considering the collective impact of all attributes, thereby ensuring that the company can identify profitable stones with greater confidence.

Code

```r
# Omit rows with missing values
train_data <- na.omit(train_data)
test_data <- na.omit(test_data)


# Fit Random Forest Model
rf_model <- randomForest(price ~ ., data = train_data, ntree = 100)

# Predict on test data
rf_predictions <- predict(rf_model, newdata = test_data)

# Calculate RMSE for Random Forest
rf_rmse <- sqrt(mean((test_data$price - rf_predictions)^2))
print(paste("RMSE for Random Forest:", rf_rmse))
```

Output

```
                Length Class  Mode
call               4  -none- call
type               1  -none- character
predicted      21029  -none- numeric
mse              100  -none- numeric
rsq              100  -none- numeric
oob.times      21029  -none- numeric
importance        10  -none- numeric
importanceSD       0  -none- NULL
localImportance    0  -none- NULL
proximity          0  -none- NULL
ntree              1  -none- numeric
mtry               1  -none- numeric
forest            11  -none- list
coefs              0  -none- NULL
y              21029  -none- numeric
test               0  -none- NULL
inbag              0  -none- NULL
terms              3  terms  call
        IncNodePurity
X          1389343605
carat    110845257824
cut        1354919294
color      9583908351
clarity   17488218232
depth      1826584308
table      1353880945
x         77944402215
y         81903581994
z         34673763972
```

The Random Forest model output prioritizes the attributes of cubic zirconia that most significantly impact pricing. The 'carat' size has emerged as the primary predictor, followed by the dimensions of the stones ('x', 'y', 'z'), suggesting that these factors should be the focus of both production optimization and marketing strategies to enhance profitability.

While the model summary does not explicitly provide an Out-of-Bag error estimate, this measure is typically used in Random Forest models to assess prediction accuracy. A lower Out-of-Bag error would reflect well on the model's reliability, an aspect that is crucial for confident application in price setting.

The complexity of the Random Forest model, with its ensemble approach, affords a higher predictive accuracy over a single decision tree but at the expense of computational intensity and decreased interpretability. The business implications of this complexity are significant: while providing nuanced insights into pricing dynamics, the model may necessitate greater resources for maintenance and analysis.

Model 3 - XGBoost Model

The inclusion of the XGBoost (Extreme Gradient Boosting) model in our predictive analytics suite represents an advanced approach to understanding the price determinants of cubic zirconia stones. XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. It is renowned for its performance and speed in predictive accuracy.

Model Performance and Business Implications

The XGBoost model's predictive capability is typically characterized by its ability to handle a wide range of data science problems with precision. For Gem Stones Co. Ltd, the application of the XGBoost model likely provided an even higher level of predictive accuracy due to its sophisticated handling of non-linear relationships and interactions between attributes.

This model excels in identifying complex patterns within the data, which are not as easily discerned by linear models. In a business context, this means that the XGBoost model can uncover subtle nuances that influence cubic zirconia prices. These could include interactions between the cut, clarity, and carat size that may not be immediately apparent.

Code

```
# Convert to matrices, ensuring 'price' is the last column
xgb_train <- xgb.DMatrix(data = as.matrix(train_data_final[, -ncol(train_data_final)]), label = train_data_final$price)
xgb_test <- xgb.DMatrix(data = as.matrix(test_data_final[, -ncol(test_data_final)]))

# Now proceed with XGBoost training
params <- list(
  objective = "reg:squarederror",
  eta = 0.3,
  max_depth = 6
)
xgb_model <- xgboost(params = params, data = xgb_train, nrounds = 100, verbose = 0)

# Predicting
xgb_predictions <- predict(xgb_model, newdata = xgb_test)
```

Output

```
[1] "RMSE: 604.334437075673"
[1] "MAE: 300.047996018776"
[1] "R-squared: 0.9776506138477"
```

The XGBoost model has exhibited outstanding performance in predicting cubic zirconia prices with an R-squared value of 0.9776, indicating that approximately 97.76% of the variance in the stone prices is predictable from the features included in the model. This high level of accuracy is further evidenced by a relatively low Root Mean Squared Error (RMSE) of 604.33, which reflects the average deviation of the predicted prices from the actual prices. Additionally, the Mean Absolute Error (MAE) of 300.05 offers a clear representation of the average absolute error in the price predictions. These metrics underscore the model's precision and its effectiveness as a tool for helping Gem Stones Co. Ltd. identify the most profitable stones, thereby optimizing their profit share and enhancing their market positioning. The XGBoost model's ability to accurately predict prices based on the attributes of cubic zirconia stones makes it an invaluable asset in the company's analytical arsenal.

Conclusion and Consolidated Insights from Predictive Modeling

The comprehensive analysis conducted using three predictive models – Linear Regression, Random Forest, and XGBoost – has provided Gem Stones Co. Ltd with a robust framework for understanding the pricing dynamics of cubic zirconia stones. Each model brought unique insights and value to the pricing strategy and profitability optimization for the company.

The Linear Regression model demonstrated a strong linear relationship between the stone attributes and their prices, enabling clear visibility into how individual features affect the price. However, it also highlighted the potential for underpricing in the higher-value segments, suggesting a need for careful management of premium product pricing.

The Random Forest model offered a nuanced perspective with its ensemble approach, revealing the collective impact of attributes on pricing and providing a more generalized model that reduces the risk of overfitting. The insights into attribute importance from this model are instrumental in guiding production focus and marketing efforts towards the most profitable stone features.

The XGBoost model took the analysis further with its advanced machine learning capabilities, providing an exceptional degree of predictive accuracy as evidenced by the high R-squared value and low error metrics. This model's ability to identify complex, non-linear relationships between attributes presents an opportunity for strategic refinement in both manufacturing and pricing practices.

Overall, the integration of insights from these models enables Gem Stones Co. Ltd to:

- Identify the critical attributes that drive the price of cubic zirconia stones, focusing on producing and marketing stones with characteristics most associated with higher prices.
- Develop a pricing strategy that is data-driven and capable of maximizing profitability by accurately predicting which stones will be more profitable.
- Enhance the precision of their inventory stocking decisions, ensuring a product mix that aligns with consumer demand and profitability potential.
- Establish a competitive advantage in the market by leveraging advanced analytics to offer products that meet consumer expectations in terms of quality and price.

In summary, the analysis not only responds to the immediate need for accurate price prediction but also empowers Gem Stones Co. Ltd with strategic direction for future growth. By identifying the most profitable stones and understanding the intricate pricing structure, the company is well-equipped to optimize its profit margins and secure a stronghold in the competitive landscape of cubic zirconia manufacturing.

Group Contributions :

Richard Allenpaul	Found the Dataset, Analysis, Report
Sahana Raje…	Analysis, Presentation, Report
Shail Desai	Presentation, Analysis

Note: Our team fairly contributed the same amounts in every section of this project

-