AIN SHAMS UNIVERSITY

FACULTY OF ENGINEERING

i-CREDIT HOURS ENGINEERING PROGRAMS

CSE 486

BIG-Data ANALYTICS

Diamond Price Analysis and Prediction

SUBMITTED BY:

Nouran Elsayed	23P0006
Reetaj Ahmed	23P0114
Abdelmoneim Mahmoud	23P0015
Hossam Osama	23P0010
Jana Ehab	23P0105
Hassan Sherif	23P0017

Contents

1.0 PROJECT DESCRIPTION

The objective of this project is to apply the **Big Data Analytics lifecycle** on a real-world dataset related to diamond prices.

The project covers all stages studied in the course, including data understanding, preprocessing, exploratory data analysis (EDA), hypothesis testing, and predictive modeling.

The project aims to analyze the factors affecting diamond prices and build an accurate prediction model using regression-based techniques.

2.0 DATASET DESCRIPTION

The dataset used in this project contains information about diamonds and their physical and quality attributes.

Main Attributes:

- **price:** Diamond price (target variable)
- **carat:** Weight of the diamond
- **cut:** Quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- **color:** Diamond color grade (D–J)
- **clarity:** Diamond clarity grade
- **depth:** Total depth percentage
- **table:** Width of the top of the diamond
- **length.mm., width.mm., height.mm.:** Physical dimensions
- **volume:** Computed attribute (length × width × height)

The dataset contains numerical and categorical variables, making it suitable for statistical analysis and regression modeling.

3.0 PROBLEM DEFINITION AND OBJECTIVES

Problem Definition

Diamond prices depend on multiple physical and quality-related attributes. The challenge is to identify which factors significantly affect price and build a predictive model that accurately estimates diamond prices.

Project Objectives

- Understand relationships between diamond attributes and price
- Clean and preprocess the dataset
- Perform hypothesis testing to validate attribute importance
- Apply EDA for pattern discovery
- Build and compare regression models
- Select the best predictive model

4.0 Data Visualization and Exploratory Data Analysis (Before Cleaning)

This section presents visual exploration of the **raw diamonds dataset** prior to any cleaning or preprocessing steps. The goal is to understand data distributions, detect anomalies, and examine relationships between attributes and price.

4.1 Missing Value Analysis

Visualization:

```
> print(na_counts)
        X     carat       cut     color   clarity     depth     table
        0         0         0         0         0         0         0
    price length.mm.  width.mm. height.mm.
        0         0         0         0
```

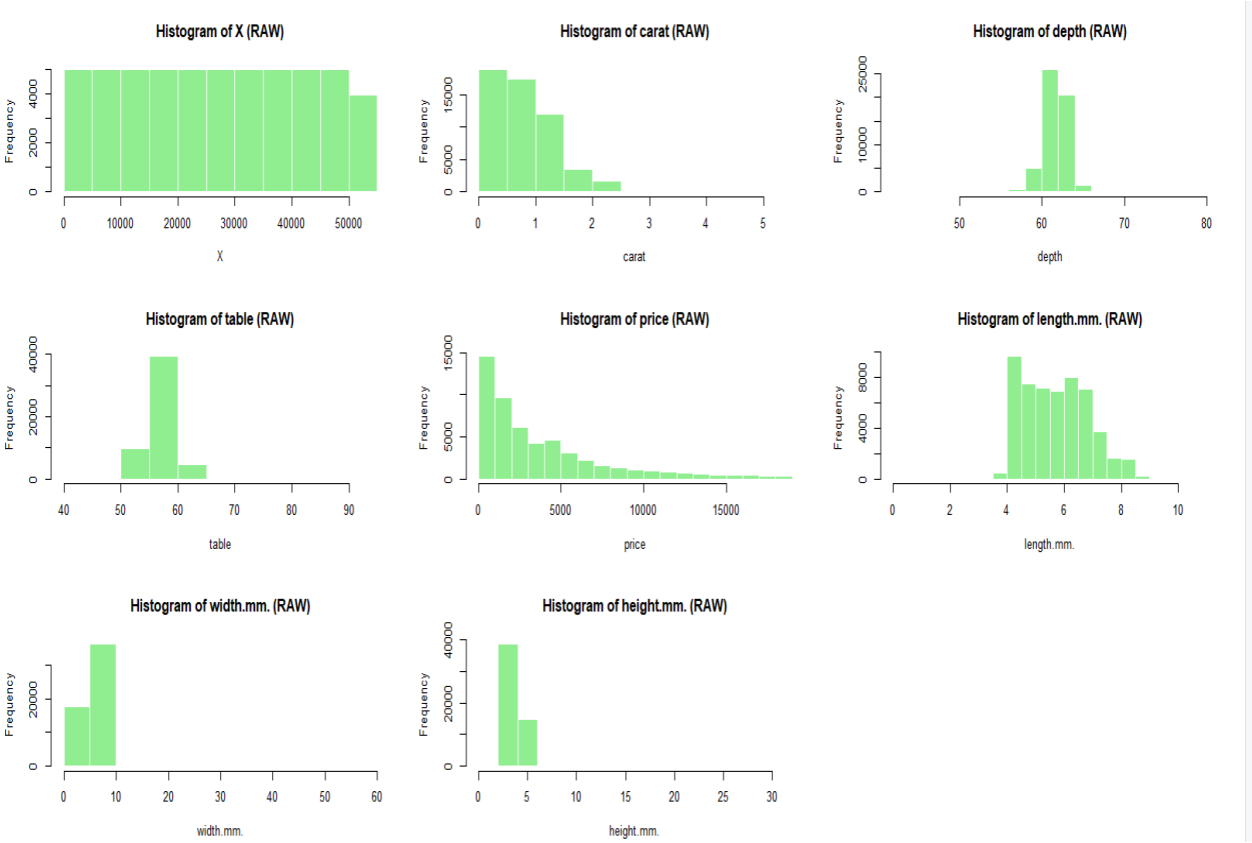
Observation & Interpretation:

The missing value check shows that **no attributes contain missing values** in the raw dataset. All numerical and categorical variables are fully populated, indicating that missing data imputation

was not required. This allowed the analysis to focus on outlier detection and data consistency rather than handling incomplete records.

4.2 Distribution of Numerical Attributes (Histograms – RAW Data)

Visualizations:



Carat

- The carat distribution is **right-skewed**, with most diamonds having low carat values.
- A small number of diamonds have very large carat values, indicating potential outliers.

Depth

- Depth values are tightly clustered around a central range.
- The distribution appears approximately normal with minor extreme values.

Table

- Table percentages are concentrated around typical industry values.

- Few extreme values appear on both ends, suggesting possible anomalies.

Price

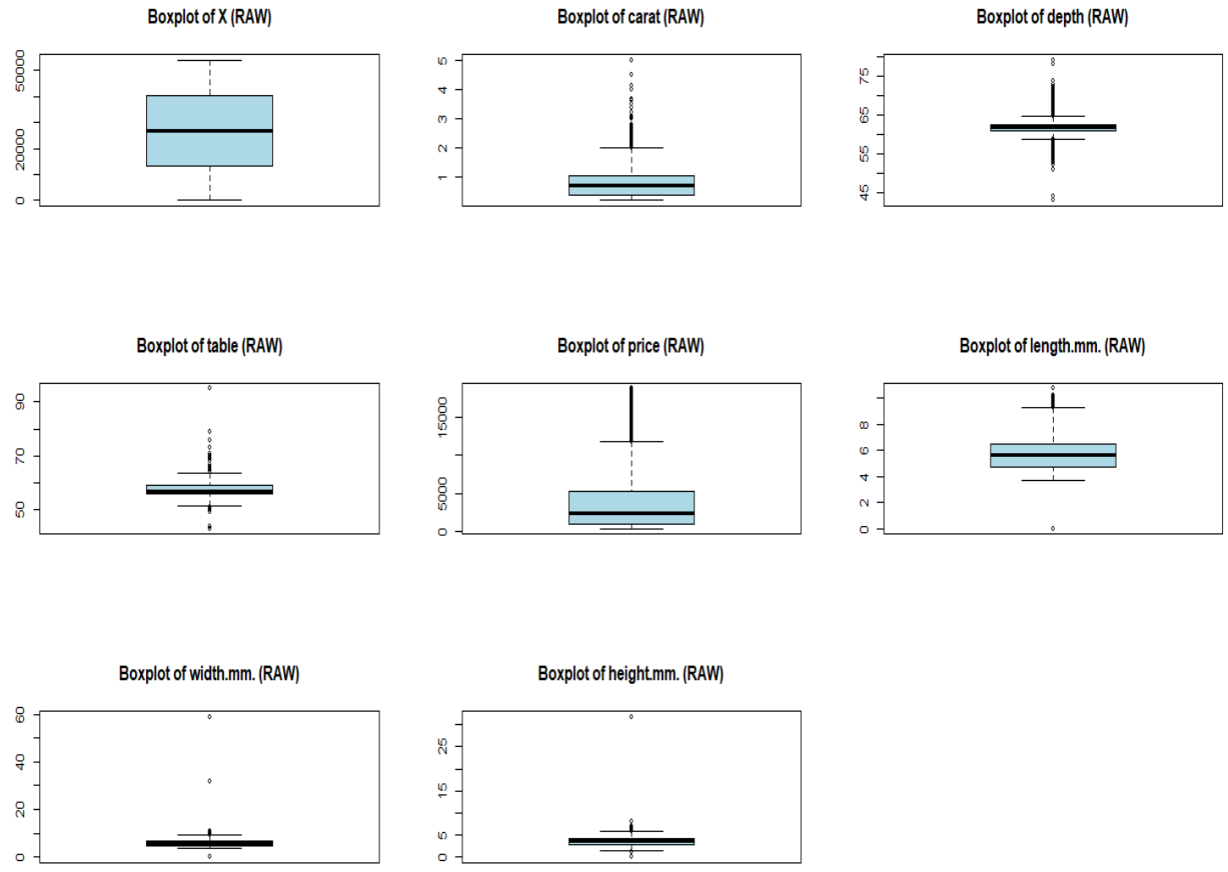
- Price shows a **highly right-skewed distribution**.
- Most diamonds are low-priced, while a small number of very expensive diamonds create a long tail.

Length, Width, and Height

- Physical dimensions show clustered distributions.
- However, extreme values (especially very large widths and heights) are visible, indicating potential invalid measurements.

4.3 Outlier Detection Using Boxplots (RAW Data)

Visualizations:

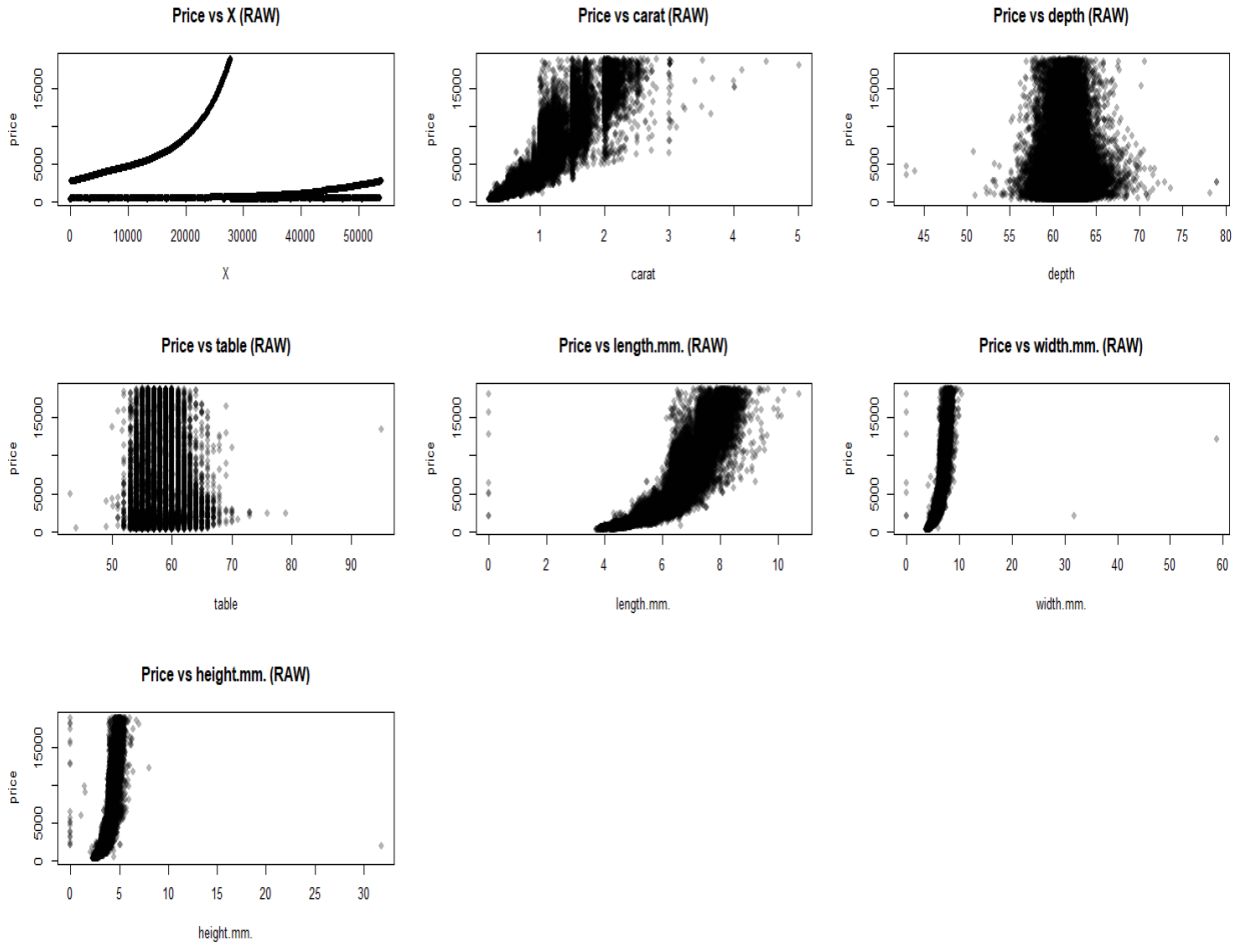


Observation & Interpretation:

- **Carat and Price** exhibit a large number of extreme values, confirming the presence of significant outliers.
- **Length, width, and height** show unusually large values far beyond the interquartile range, suggesting physically impossible diamond dimensions.
- **Depth and table** contain fewer outliers but still display some extreme observations.
- The presence of many outliers justifies the need for systematic outlier removal in later preprocessing steps.

4.4 Relationship Between Price and Numerical Attributes (Scatter Plots – RAW Data)

Visualizations:



Price vs Carat

- A **strong positive nonlinear relationship** is observed.
- Price increases rapidly as carat increases, indicating carat as a key price determinant.

Price vs Physical Dimensions

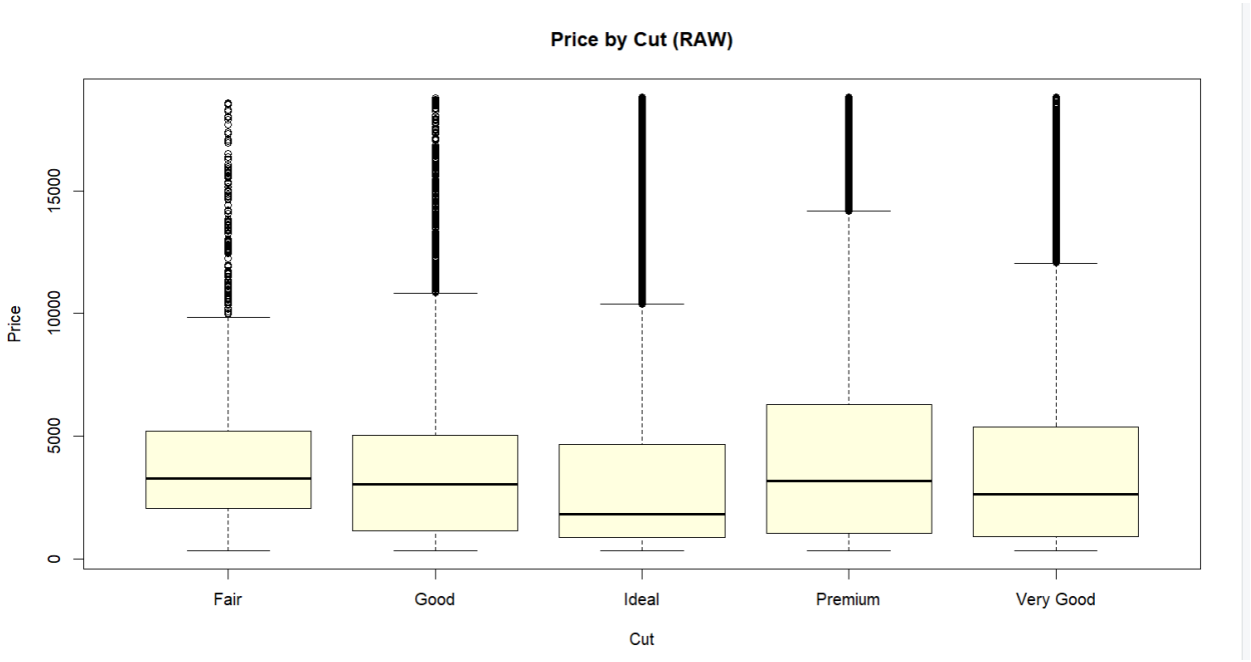
- Length, width, and height show strong positive relationships with price.
- Vertical bands and extreme points indicate measurement anomalies and outliers.

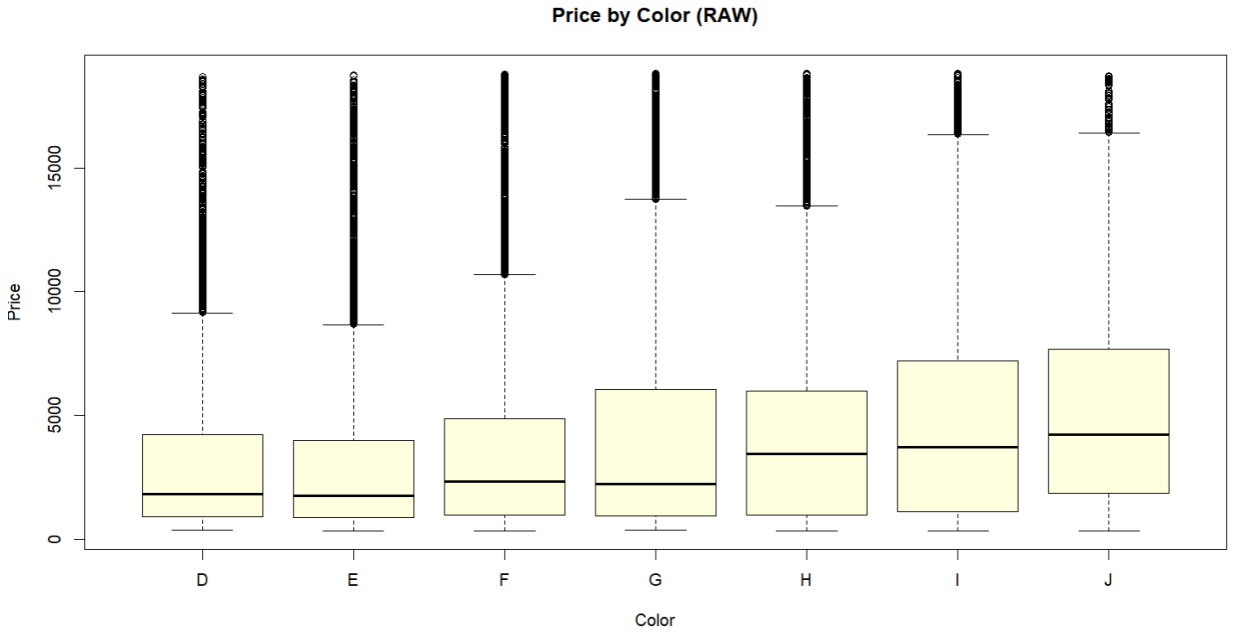
Price vs Depth and Table

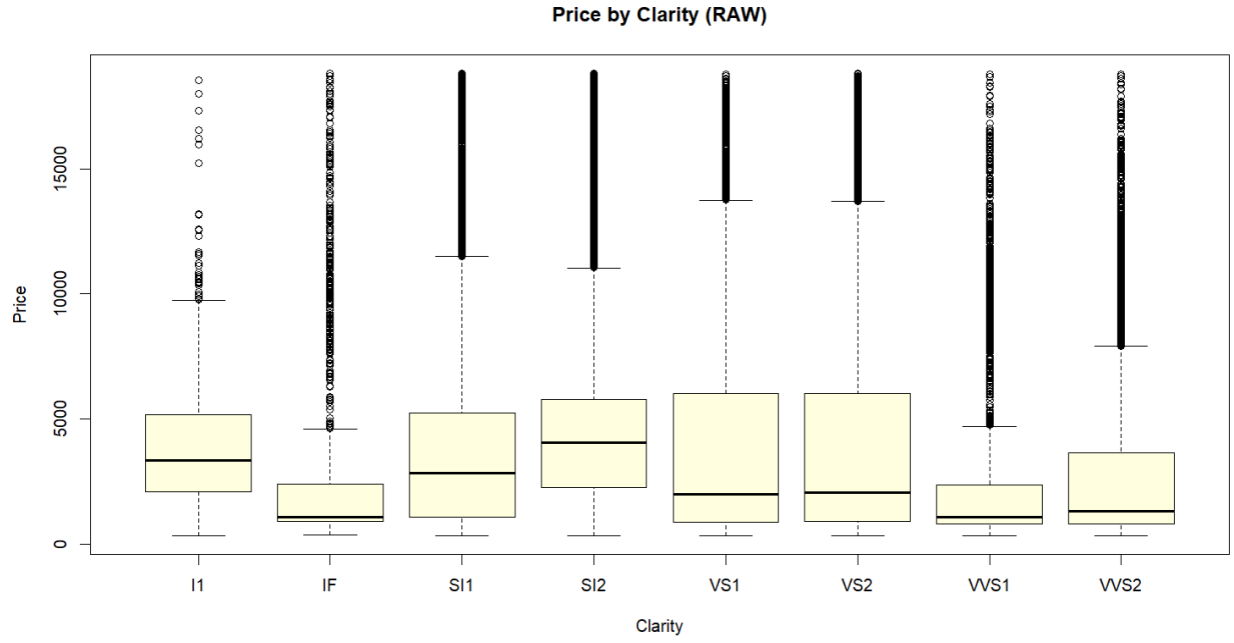
- No clear linear relationship is observed.
- The wide spread suggests depth and table alone are weak predictors of price.

4.5 Price Distribution Across Categorical Variables (RAW Data)

Visualizations:







Price by Cut

- Median prices differ across cut categories.
- Premium and Ideal cuts show wider price ranges, partly due to carat differences.

Price by Color

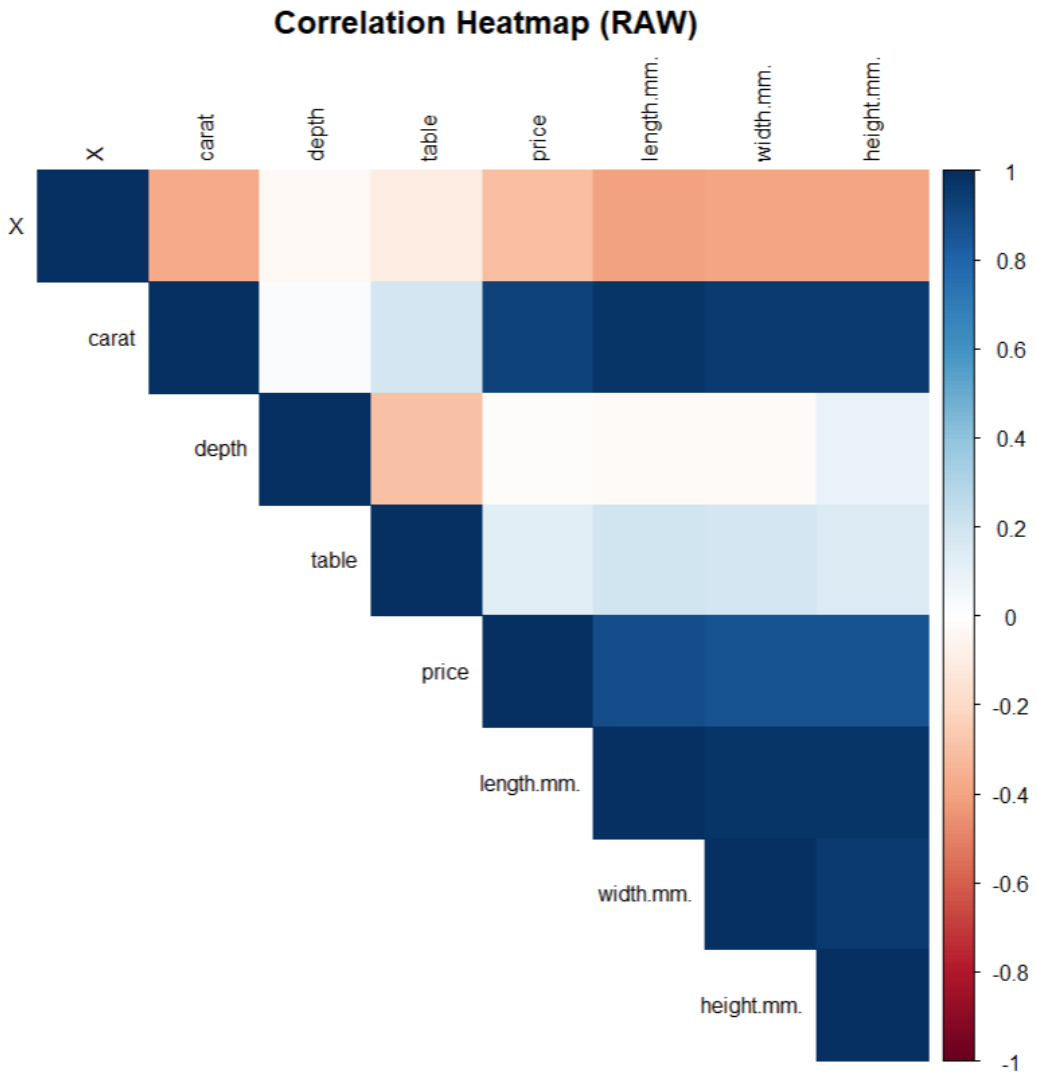
- Diamonds with lower color grades tend to have higher median prices.
- This counterintuitive trend suggests color interacts with other attributes such as carat and clarity.

Price by Clarity

- Higher clarity grades generally correspond to higher prices.
- Significant overlap exists between categories, indicating that clarity alone does not fully determine price.

4.6 Correlation Analysis (RAW Data)

Visualization:



Observation & Interpretation:

- **Price has strong positive correlations** with carat and physical dimensions.
- Depth and table show weak or near-zero correlation with price.
- High correlations among length, width, and height indicate multicollinearity, motivating the later creation of a derived volume feature.

4.7 Summary of Pre-Cleaning Insights

- The raw dataset contains **significant outliers**, especially in price, carat, and physical dimensions.
- Several physically unrealistic measurements are present.
- Strong relationships between price and size-related attributes are evident.
- These findings motivated the data cleaning, outlier removal, and feature engineering steps applied in the preprocessing phase.

5.0 DATA PREPROCESSING AND CLEANING

Data preprocessing was a critical step in this project to ensure the quality, reliability, and consistency of the dataset before performing exploratory analysis, hypothesis testing, and predictive modeling. The raw dataset (Diamonds Prices 2022) was first examined to understand its structure, data types, and summary statistics.

5.1 Initial Data Inspection

The dataset was inspected to identify:

- The number of observations and attributes
- Data types of each attribute
- Possible data quality issues such as extreme values or invalid entries

This initial inspection helped guide the subsequent cleaning steps.

5.2 Outlier Detection

Outliers were identified in all numerical attributes using the **Interquartile Range (IQR) method**. For each numeric feature, the first quartile (Q1), third quartile (Q3), and IQR were calculated. Any values lying outside the range:

$$[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$$

were considered outliers.

The number of lower, upper, and total outliers was recorded for each numeric attribute to understand the extent of extreme values present in the dataset.

```
> outlier_counts
      Column Lower_Outliers Upper_Outliers Total_Outliers
1          X              0              0              0
2      carat              0           1889           1889
3      depth           1506           1039           2545
4      table             16            589            605
5      price              0           3540           3540
6 length.mm.              8             24             32
7  width.mm.              7             22             29
8 height.mm.             21             28             49
>
```

5.3 Outlier Removal

After identifying outliers, rows containing extreme values were removed from the dataset. This step reduced noise and prevented extreme observations from disproportionately influencing statistical analysis and regression models.

The number of records before and after outlier removal was compared to ensure that a substantial portion of the data was preserved while improving overall data quality.

```
> cat("Original rows:", nrow(df), "\n")
Original rows: 53943
> cat("Rows after removing outliers:", nrow(df_clean), "\n")
Rows after removing outliers: 46535
```

5.4 Removal of Invalid and Impossible Values

Additional logical constraints were applied to ensure all records represented valid diamonds:

- Diamonds with non-positive **carat weight** or **price** were removed.
- **Depth percentage** was restricted to a realistic range.
- Records with zero or negative physical dimensions (length, width, or height) were excluded.

These filters eliminated physically impossible or unreliable observations.

5.5 Handling Missing Values

A missing-value check was performed after cleaning. The final dataset contained **no missing values**, ensuring completeness for statistical tests and model training.

```
> df_clean <- df_clean[df_clean$carat > 0, ]
> df_clean <- df_clean[df_clean$price > 0, ]
> df_clean <- df_clean[df_clean$depth >= 40 & df_clean$depth <= 80, ]
> df_clean <- df_clean[df_clean$length.mm. > 0, ]
> df_clean <- df_clean[df_clean$width.mm. > 0, ]
> df_clean <- df_clean[df_clean$height.mm. > 0, ]
>
> colSums(is.na(df_clean))
         X      carat        cut      color    clarity      depth
         0          0          0          0          0          0
     table      price length.mm.  width.mm. height.mm.
         0          0          0          0          0
```

5.6 Data Type Correction

Categorical attributes such as cut, color, and clarity were converted to categorical data types (factors).
This conversion was necessary for:

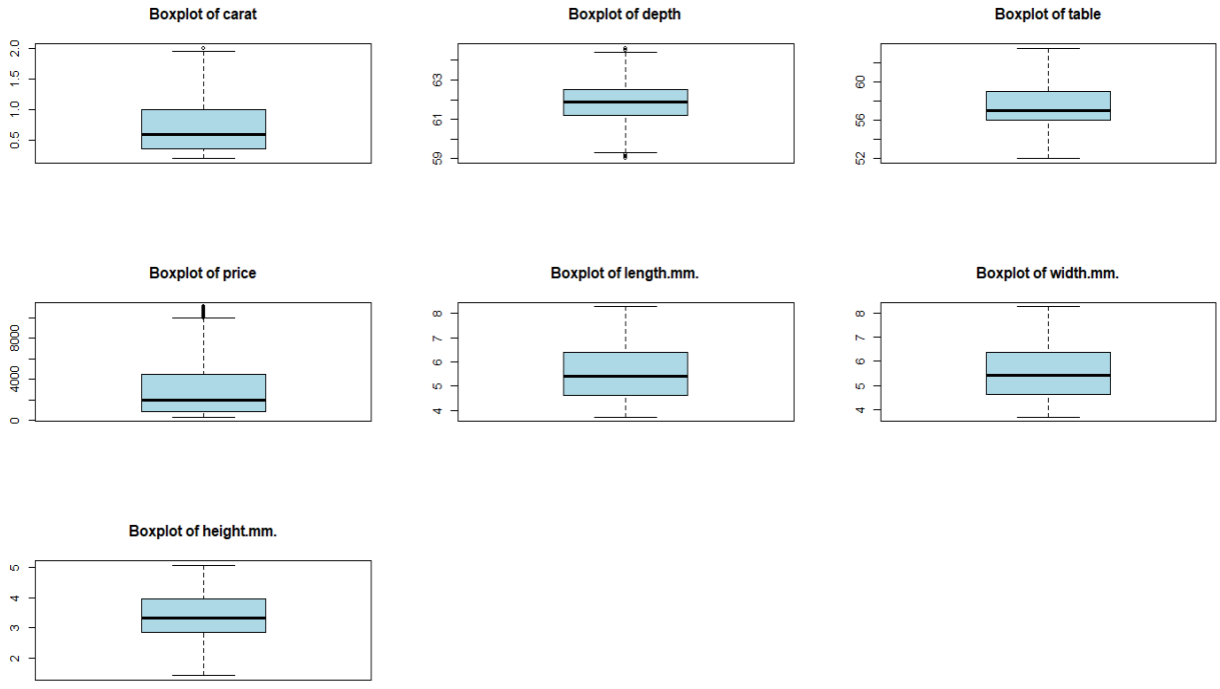
- Analysis of variance (ANOVA)
- Regression models involving categorical predictors

5.7 Duplicate Removal and Column Refinement

Duplicate records were removed to prevent redundancy and biased results. Additionally, an unnecessary index column was removed to keep the dataset clean and meaningful.

5.8 Outlier Visualization

Boxplots were generated for all numerical attributes after cleaning to visually verify the distribution of values and confirm the effectiveness of outlier removal.



- **Carat:** The distribution is right-skewed, with a small number of higher-value observations near the upper whisker. These represent naturally larger diamonds rather than data errors.
- **Depth:** Values are tightly clustered around the median, with very limited variability, indicating no significant outliers after cleaning.
- **Table:** The table percentage shows a compact distribution with no extreme outliers, suggesting consistent measurements.
- **Price:** The price distribution remains right-skewed, with some high-value observations. These correspond to premium diamonds and are considered valid rather than anomalous.
- **Length, Width, and Height:** All physical dimension attributes exhibit stable distributions with no abnormal values, confirming the removal of invalid or impossible measurements.

Overall, the boxplots confirm that the applied outlier detection and filtering steps were effective. Remaining extreme values reflect genuine variations in diamond characteristics and were retained to preserve meaningful information for analysis and modeling.

5.9 Feature Engineering

A new attribute, volume, was created using the diamond's physical dimensions (length, width, and height).

This feature represents the actual physical size of the diamond and was later found to be a strong predictor of price.

5.10 Final Cleaned Dataset

After completing all preprocessing steps, the final cleaned dataset was saved as **Diamonds_Final_Cleaned.csv**.
This dataset served as the input for:

- Exploratory Data Analysis (EDA)
- Hypothesis testing
- Predictive modeling

6.0 HYPOTHESIS TESTING AND STATISTICAL ANALYSIS

This section presents statistical hypothesis testing conducted on the **cleaned diamonds dataset** to examine the relationship between diamond price and its numerical and categorical attributes. Pearson correlation tests were used for numerical variables, while Analysis of Variance (ANOVA) and post-hoc Tukey tests were applied to categorical variables.

6.1 Correlation Analysis for Numerical Attributes

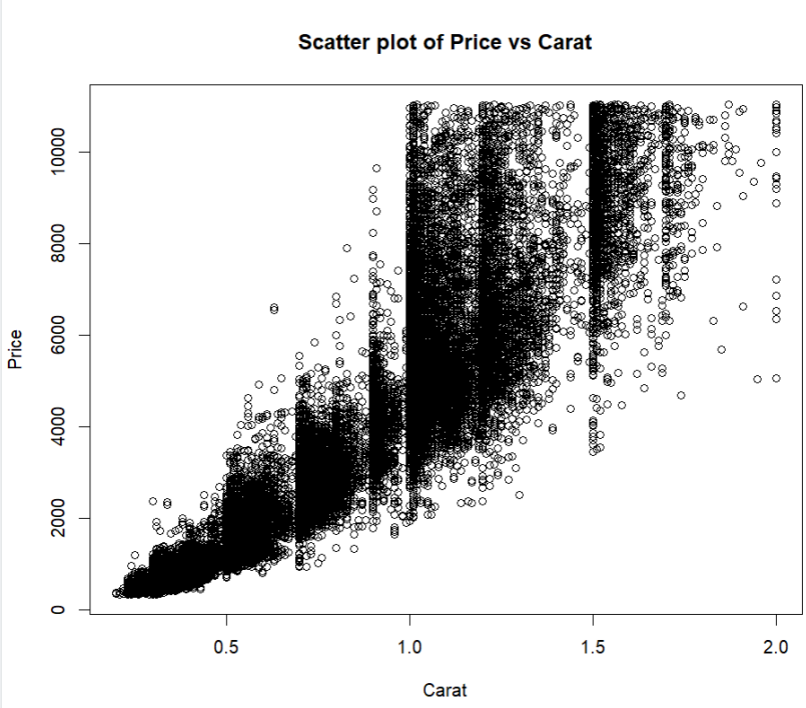
Pearson's correlation test was applied to evaluate the linear relationship between diamond price and selected numerical features.

- **Price vs Carat**

  The correlation analysis revealed a very strong positive relationship between carat weight and price
  ($r = 0.9248, p < 2.2 \times 10^{-16}$).

  This indicates that as the carat weight increases, diamond price increases significantly.

  The narrow confidence interval further confirms the stability of this relationship. The scatter plot shows a clear upward nonlinear trend, emphasizing carat as the most influential numerical predictor of price.

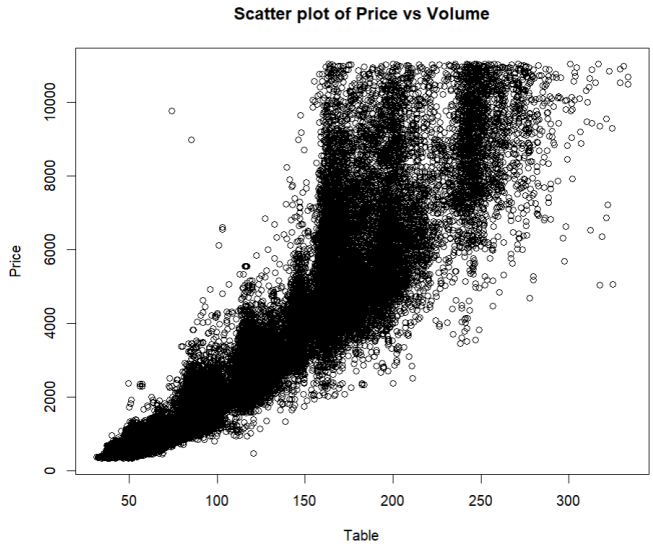


- **Price vs Volume**

The derived volume feature also demonstrated a strong positive correlation with price ($r = 0.9259, p < 2.2 \times 10^{-16}$).

This confirms that overall physical size, represented by volume, is a key determinant of diamond price.

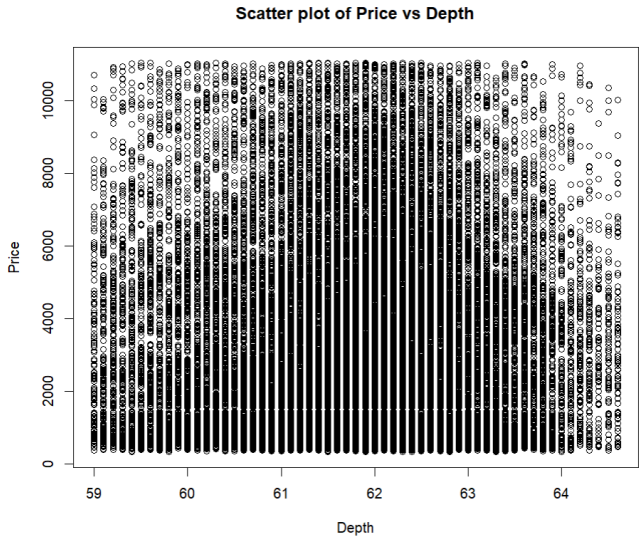
The correlation strength is comparable to carat, validating the decision to include volume as an engineered feature in predictive modeling.



- **Price vs Depth**

The relationship between depth and price was found to be extremely weak ($r = 0.0144$), despite being statistically significant ($p = 0.00197$).

This statistical significance is largely due to the large sample size rather than a meaningful effect. The scatter plot shows no visible trend, indicating that depth contributes negligibly to price variation and was therefore excluded from prediction models.

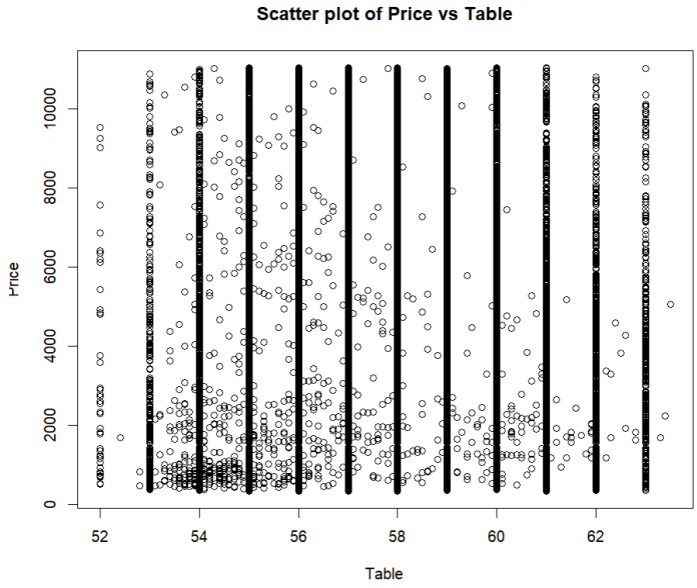


- **Price vs Table**

The table percentage exhibited a weak positive correlation with price ($r = 0.1401$, $p < 2.2 \times 10^{-16}$).

Although statistically significant, the low correlation coefficient and scattered pattern indicate that table alone is a poor predictor of price and was not prioritized in modeling.

.



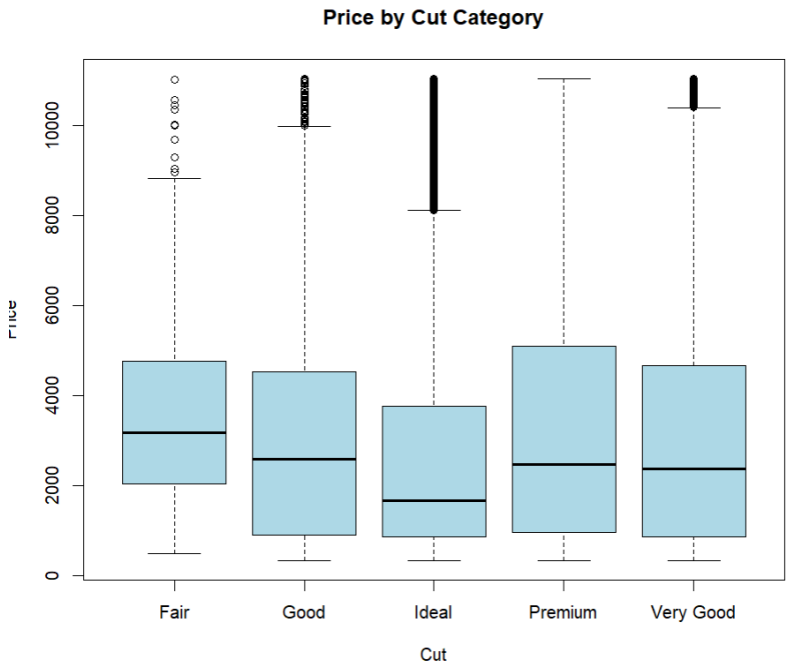
6.2 Analysis of Variance (ANOVA) for Categorical Attributes

ANOVA tests were conducted to examine whether diamond price differs significantly across categories of cut, color, and clarity.

Effect of Cut on Price

The ANOVA results indicate a **statistically significant difference** in mean prices among cut categories
($F = 130, p < 2 \times 10^{-16}$).

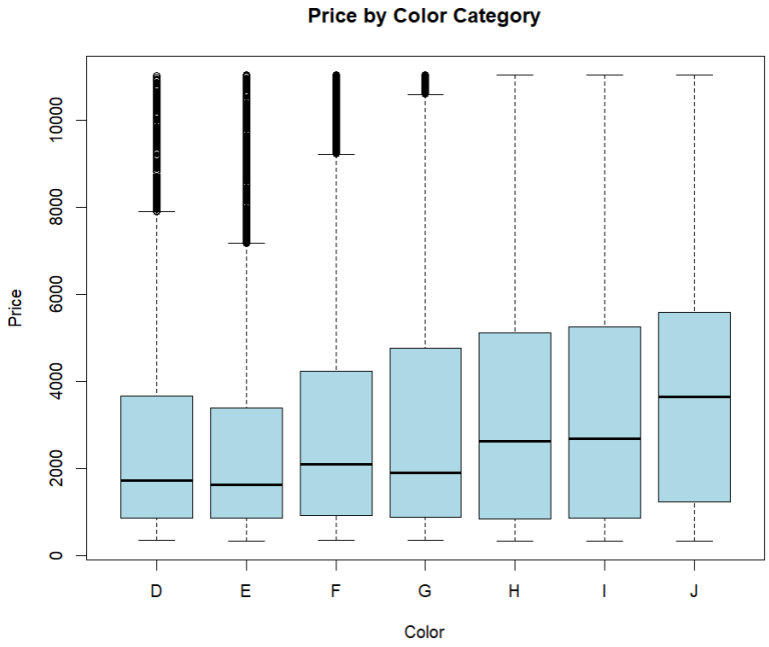
Tukey's post-hoc test showed that most cut categories differ significantly in price. However, some pairs such as **Very Good vs Good** and **Premium vs Fair** did not show statistically significant differences, indicating overlapping price ranges between adjacent quality levels.



Effect of Color on Price

The ANOVA test for color revealed a **strong and statistically significant effect** on price ($F = 178.5, p < 2 \times 10^{-16}$).

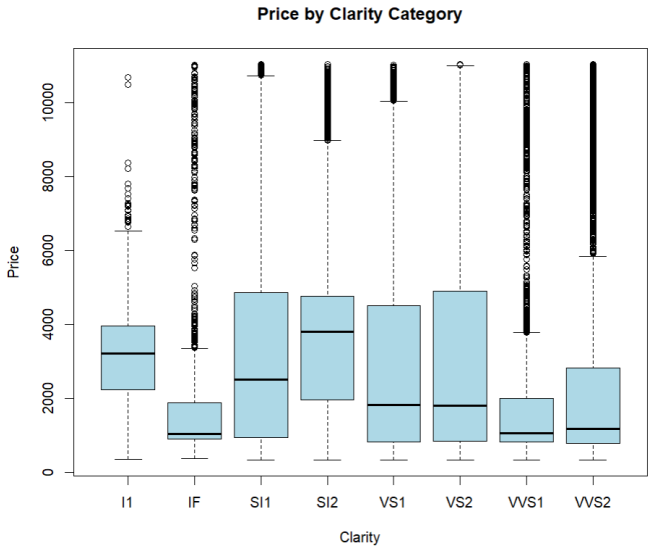
Tukey's HSD test confirmed that most color categories differ significantly from one another. Diamonds with lower color grades (H–J) generally exhibited higher prices, largely due to interaction effects with carat and clarity.



Effect of Clarity on Price

Clarity showed the **strongest categorical impact** on price ($F = 212.1, p < 2 \times 10^{-16}$).

Post-hoc analysis revealed significant price differences across most clarity grades. Higher clarity levels (VVS1, VVS2, IF) generally corresponded to higher prices, though some adjacent categories showed overlapping price distributions.



6.3 Summary of Hypothesis Testing Results

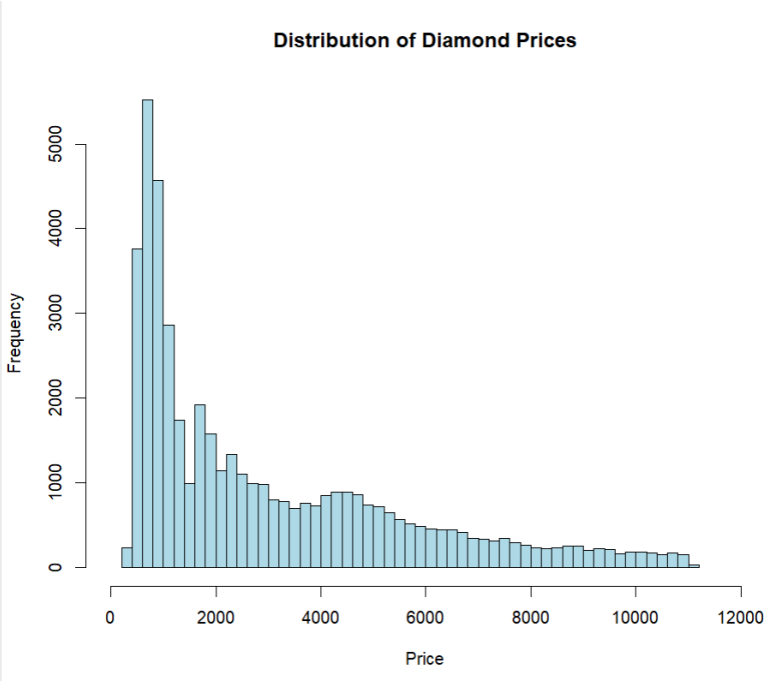
- **Carat and volume** have strong positive relationships with price.
- **Depth and table** exhibit weak relationships and contribute little explanatory power.
- **Cut, color, and clarity** all show statistically significant differences in price distributions.
- These findings guided **feature selection** for the predictive modeling phase.

7.0 EXPLORATORY DATA ANALYSIS(EDA) AFTER CLEANING

This section presents the exploratory data analysis performed after applying data cleaning steps. The objective is to understand the distributions of key variables and their relationships with diamond price using visualizations. Screenshots of the corresponding graphs are provided separately.

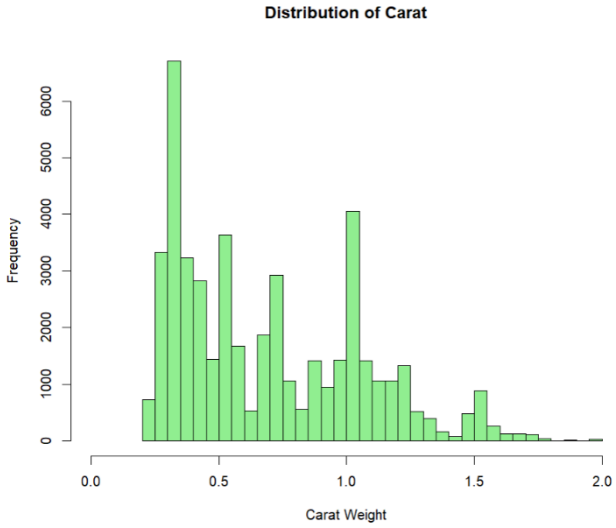
7.1 Distribution of Diamond Prices

The histogram of diamond prices shows a right-skewed distribution. Most diamonds are concentrated in the lower price range, while a smaller number of diamonds have very high prices. This indicates that expensive diamonds are relatively rare, which is expected in real market data.



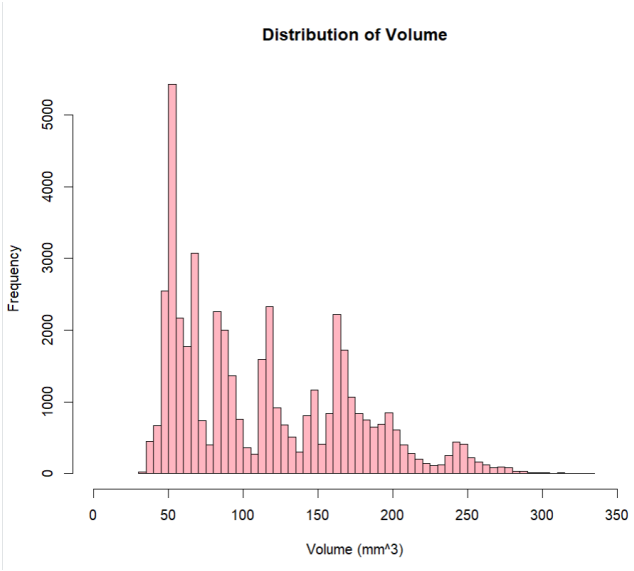
7.2 Distribution of Carat

The carat distribution is also right skewed, with the majority of diamonds having smaller carat weights. As carat increases, the frequency decreases significantly, reflecting that large diamonds are less common.



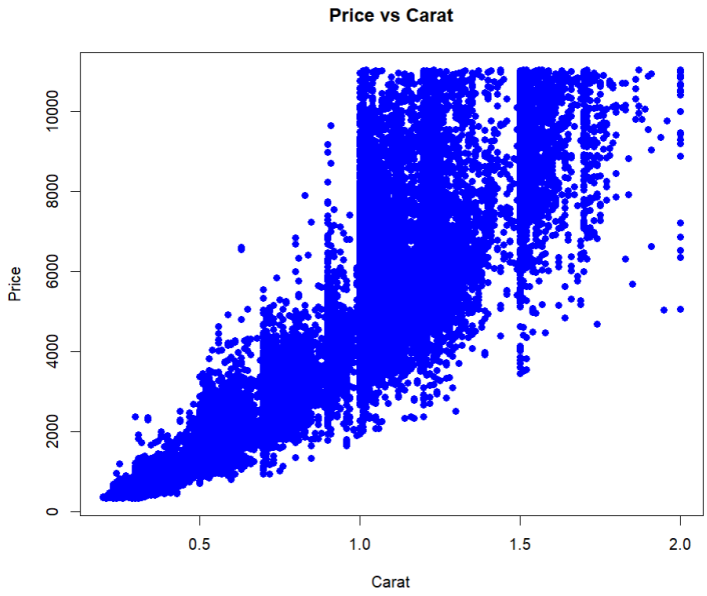
7.3 Distribution of Volume

The volume histogram shows a **positively skewed distribution** with most diamonds clustered at lower volume values. A long tail is observed for higher volumes, corresponding to larger diamonds. This confirms that volume varies widely but is dominated by smaller stones.



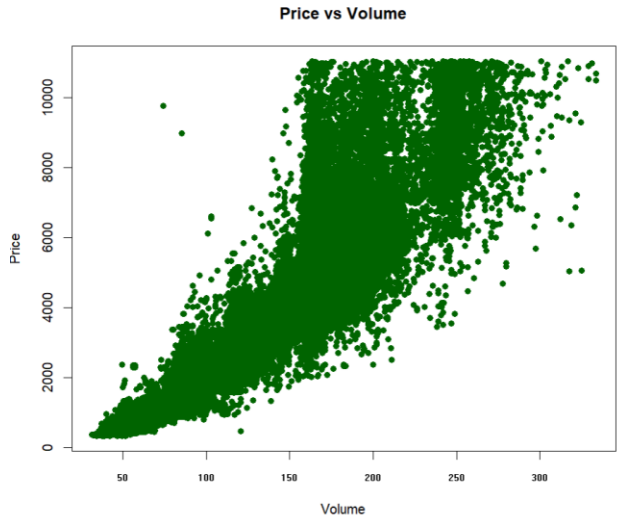
7.4 Price vs Carat

The scatter plot of price versus carat reveals a **strong positive relationship**. As carat weight increases, diamond price increases sharply. The pattern is non-linear, indicating that price grows faster for larger carat sizes. This supports the statistical correlation results.



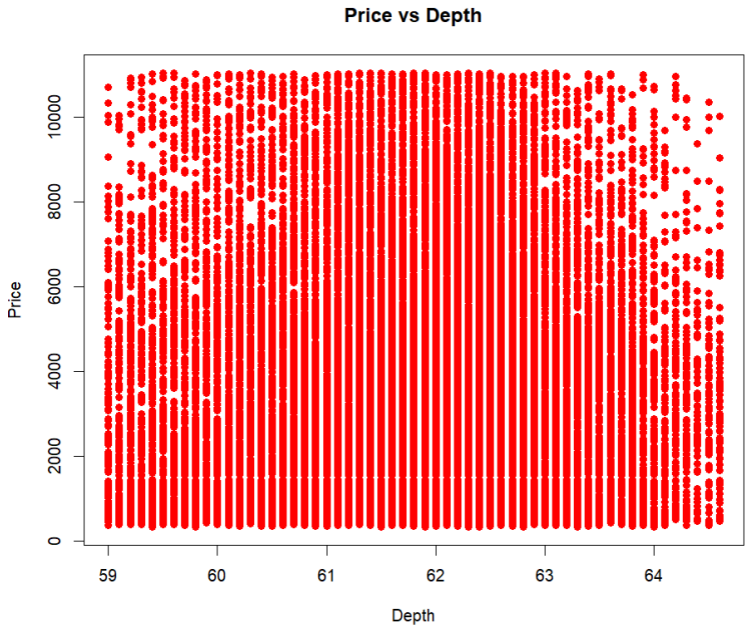
7.5 Price vs Volume

A strong **positive association** is observed between price and volume. Diamonds with larger physical size tend to have higher prices. The spread increases for larger volumes, suggesting greater price variability among bigger diamonds.



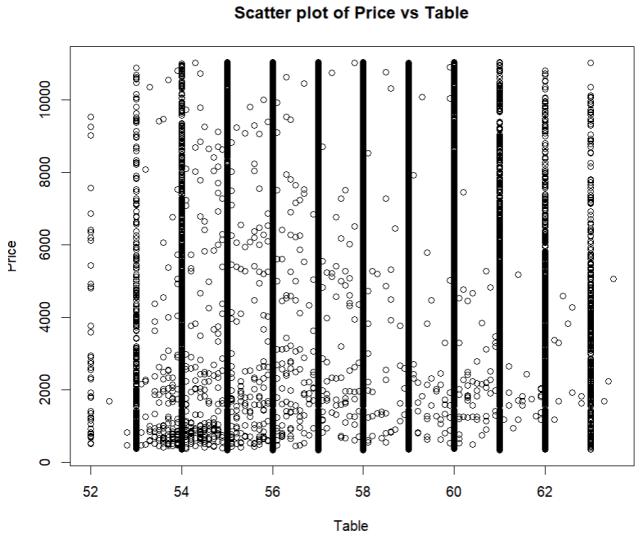
7.6 Price vs Depth

The scatter plot of price versus depth shows **no clear relationship**. Prices are widely scattered across the depth range, indicating that depth alone has little influence on price. This visual observation aligns with the very weak correlation obtained in numerical tests.



7.7 Price vs Table

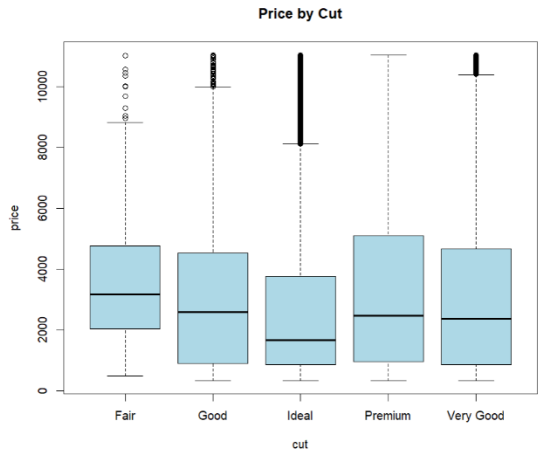
The price versus table plot shows a **weak and dispersed pattern**. Although some variation exists, there is no strong trend, suggesting that table percentage has limited impact on diamond price compared to size-related features.



7.8 Price by Cut

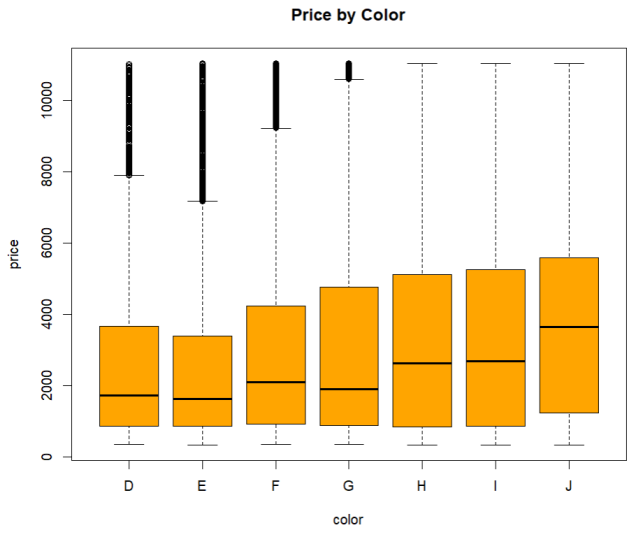
The boxplot of price by cut category shows noticeable differences between cut types.

- *Premium* and *Very Good* cuts generally exhibit higher median prices.
- *Ideal* cut diamonds show a lower median price, likely due to smaller average carat sizes. Overall, cut quality influences price, but its effect is intertwined with carat weight.



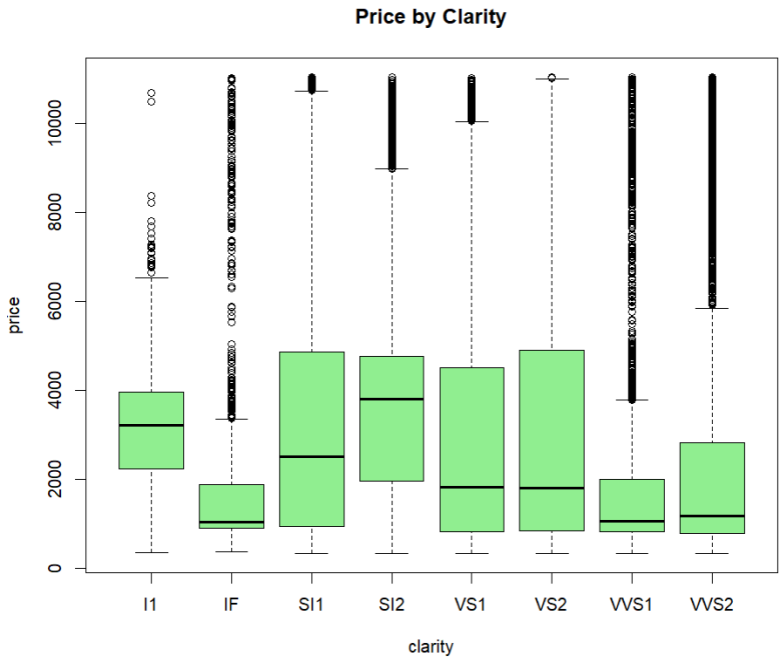
7.9 Price by Color

The boxplot indicates that diamond price generally **increases as color quality decreases** (from D to J). Lower color grades tend to have higher prices because they are often associated with larger carat sizes. Significant variation exists within each color category.



7.10 Price by Clarity

The clarity boxplot shows substantial differences in price distributions across clarity levels. Diamonds with lower clarity grades (e.g., SI and VS categories) often have higher prices due to larger sizes, while higher clarity grades (VVS, IF) show lower median prices. This highlights the combined effect of clarity and carat on pricing.



7.11 Correlation Matrix of Numeric Features (After Cleaning)

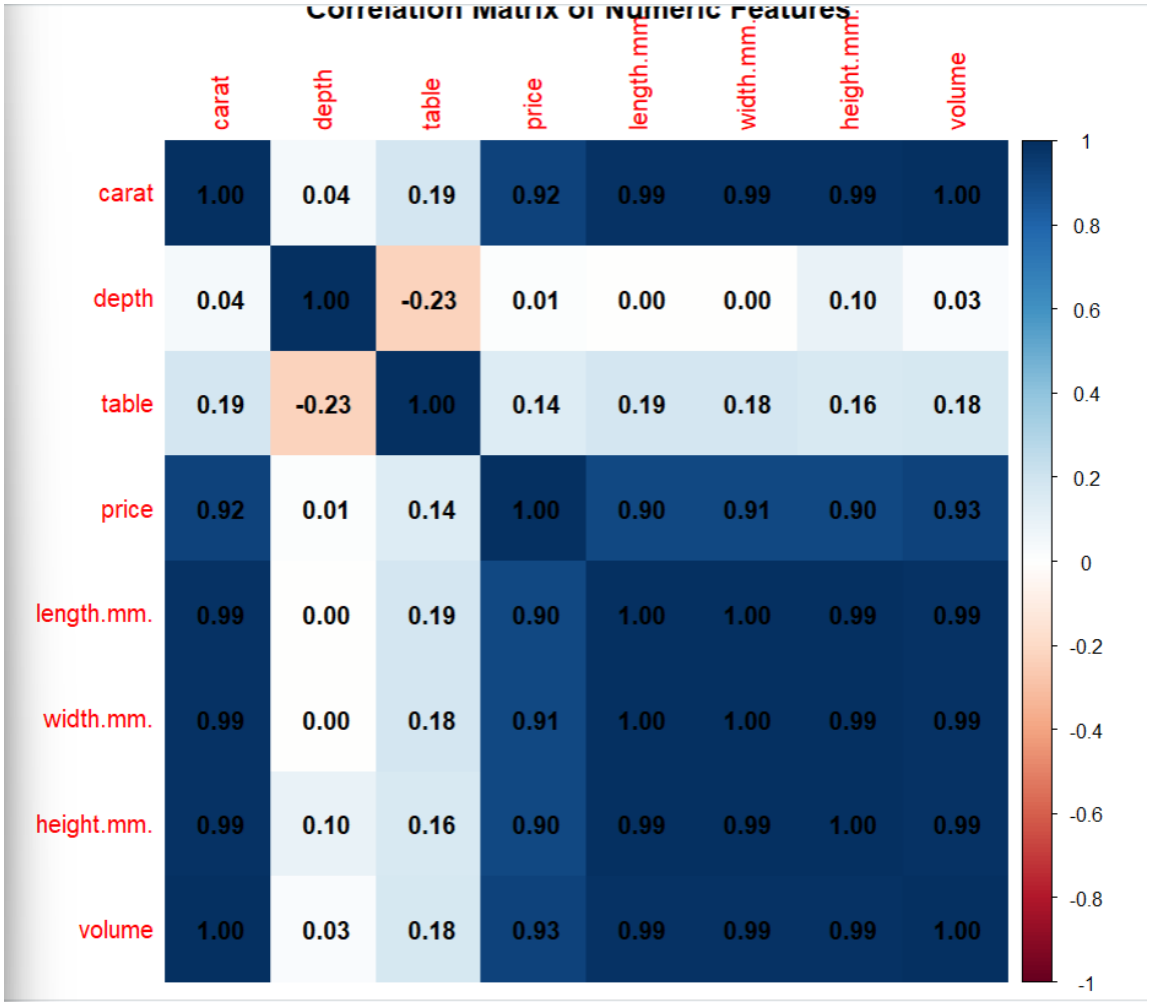
The correlation heatmap illustrates the strength and direction of linear relationships among the numeric variables in the cleaned dataset.

- Price shows a very strong positive correlation with carat (0.92), volume (0.93), and physical dimensions (length, width, and height ≈ 0.90–0.91). This confirms that diamond price is primarily driven by size-related attributes.
- Carat and volume are almost perfectly correlated (≈ 1.00), indicating that volume is largely determined by carat and the physical dimensions of the diamond.
- The physical dimensions (length, width, height) are highly correlated with each other (≈ 0.99–1.00), which is expected since they jointly describe the diamond's size.
- Depth has very weak correlation with price (≈ 0.01) and with most size-related variables, suggesting minimal influence on pricing.

- Table shows only weak correlations with price and other features (≈ 0.14–0.19), indicating limited predictive importance.

Conclusion:

The correlation analysis highlights that size-related features (carat, volume, and dimensions) are the most influential predictors of diamond price, while depth and table contribute little explanatory power. These results guided the feature selection process in the modeling stage.



8.0 DATASET PREPARATION FOR MACHINE LEARNING

8.1 Dataset Preparation and Splitting

After completing data cleaning, preprocessing, and exploratory data analysis, the final cleaned dataset (Diamonds_Final_Cleaned.csv) was prepared for machine learning modeling. The dataset contains both numerical and categorical attributes describing the physical and quality characteristics of diamonds, with **price** as the target variable to be predicted.

To build and evaluate supervised learning models, the dataset was divided into two subsets:

- **Training set (80%)**: Used to train the machine learning models.
- **Testing set (20%)**: Used to evaluate model performance on unseen data.

The split was performed using a stratified sampling approach to preserve the distribution of the target variable. This ensures that both training and testing sets are representative of the original dataset.

8.2 Feature Selection and Target Variable

- **Target Variable**:
    - price (continuous variable representing diamond price)
- **Input Features**:
    - **Numerical features**:
      carat, volume
    - **Categorical features**:
      cut, color, clarity

Categorical variables were converted into factor variables to allow proper handling by regression and tree-based algorithms. In addition, a new numerical feature (**volume**) was derived from the diamond's physical dimensions (length × width × height) to capture overall size more accurately than individual dimensions.

8.3 Reproducibility and Consistency

To ensure reproducibility of the experimental results, a fixed random seed was applied before splitting the dataset. This guarantees that the same training and testing partitions can be recreated, enabling consistent comparison across different modeling techniques.

Outcome:
This dataset preparation resulted in a well-structured training and testing framework suitable for supervised machine learning, allowing fair and reliable evaluation of multiple predictive models.

9.0 DATA ANALYTICS TECHNIQUES USED AND JUSTIFICATION

9.1 Overview of Applied Techniques

In this project, multiple **supervised learning techniques** were applied to predict diamond prices based on their physical and quality characteristics. Since the target variable (**price**) is continuous, **regression-based models** were selected. Using more than one technique allows performance comparison and strengthens the reliability of the findings.

The following data analytics techniques were used:

- **Multiple Linear Regression**
- **Support Vector Regression (SVR)**
- **Decision Tree Regression**

9.2 Multiple Linear Regression

Multiple Linear Regression was used as a **baseline predictive model**. It models the linear relationship between the diamond price and several explanatory variables such as carat, volume, cut, color, and clarity.

Justification for choice:

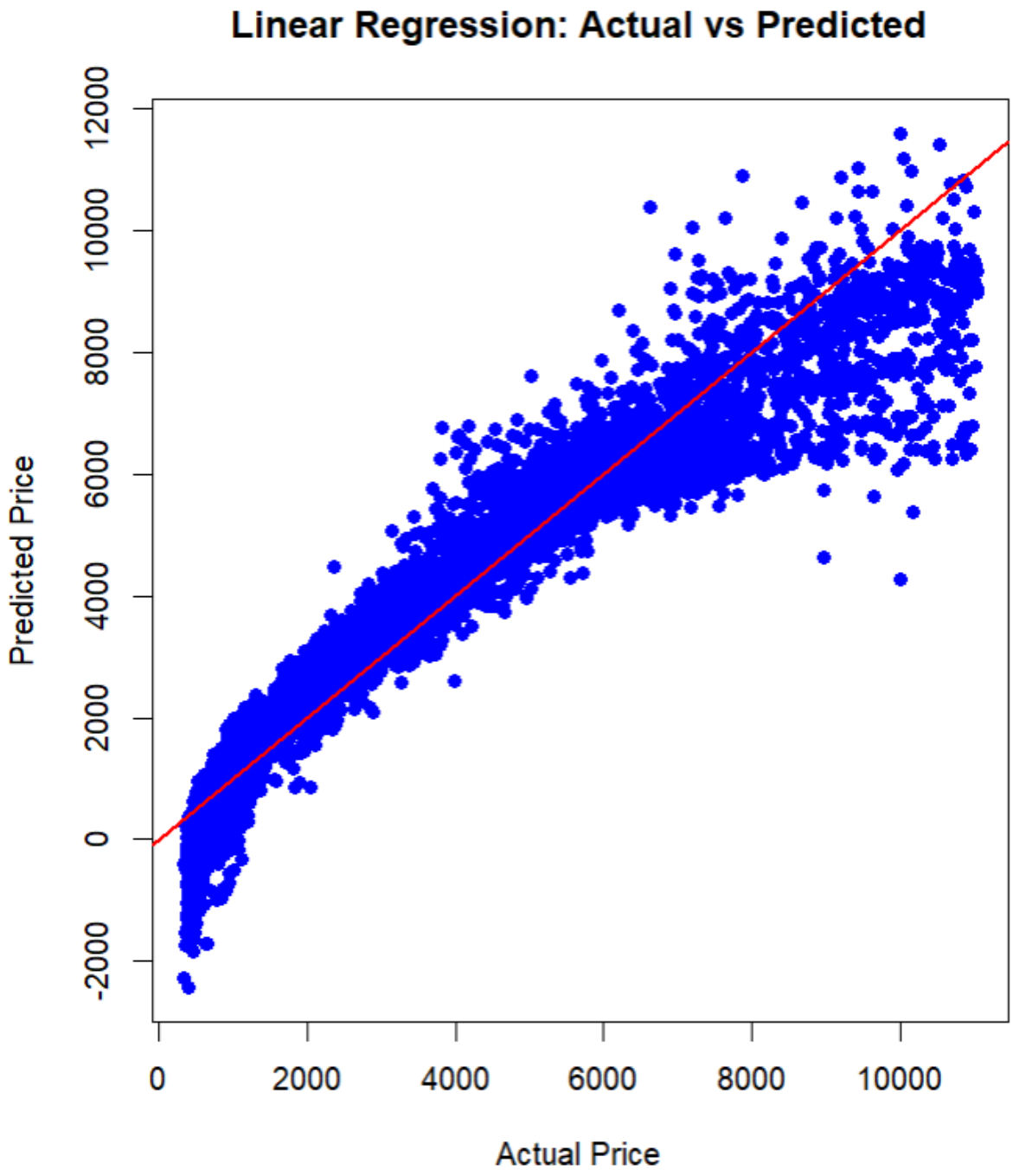
- Simple and interpretable model
- Helps understand the direct influence of each feature on price
- Serves as a benchmark to compare more complex models
- Widely used in economic and pricing analysis

This model provides insight into how diamond attributes contribute linearly to price variation.

9.2.1 Linear Regression Output

```
> summary(reg_model)

Call:
lm(formula = price ~ carat + volume + cut + color + clarity,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4634.6  -448.9  -115.0   306.3  4733.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4882.437     60.032 -81.330  < 2e-16 ***
carat        3820.140    245.209  15.579  < 2e-16 ***
volume         23.514      1.514  15.526  < 2e-16 ***
cutGood       240.534     49.803   4.830 1.37e-06 ***
cutIdeal      386.081     48.865   7.901 2.84e-15 ***
cutPremium    315.325     48.804   6.461 1.05e-10 ***
cutVery Good  308.334     48.879   6.308 2.86e-10 ***
colorE       -179.635     13.478 -13.328  < 2e-16 ***
colorF       -231.905     13.712 -16.912  < 2e-16 ***
colorG       -333.288     13.403 -24.866  < 2e-16 ***
colorH       -669.383     14.347 -46.658  < 2e-16 ***
colorI      -1031.426     16.290 -63.316  < 2e-16 ***
colorJ      -1648.230     20.417 -80.729  < 2e-16 ***
clarityIF    3518.849     42.861  82.099  < 2e-16 ***
claritySI1   2306.331     38.056  60.604  < 2e-16 ***
claritySI2   1620.311     38.313  42.291  < 2e-16 ***
clarityVS1   3047.857     38.664  78.829  < 2e-16 ***
clarityVS2   2783.979     38.217  72.847  < 2e-16 ***
clarityVVS1  3366.490     40.302  83.531  < 2e-16 ***
clarityVVS2  3375.038     39.537  85.363  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 715.7 on 37119 degrees of freedom
Multiple R-squared:  0.9243,    Adjusted R-squared:  0.9242
F-statistic: 2.384e+04 on 19 and 37119 DF,  p-value: < 2.2e-16
```

```
REGRESSION PERFORMANCE METRICS
 Accuracy (R²): 0.9236
 MAE          : 512.578
 MSE          : 514960.8
 RMSE         : 717.6077
```


Linear Regression: Actual vs Predicted
Predicted Price
-2000
0
2000
4000
6000
8000
10000
12000
0
2000
4000
6000
8000
10000
Actual Price

9.3 Support Vector Regression (SVR)

Support Vector Regression (SVR) with a radial kernel was applied to capture **non-linear relationships** between diamond features and price.

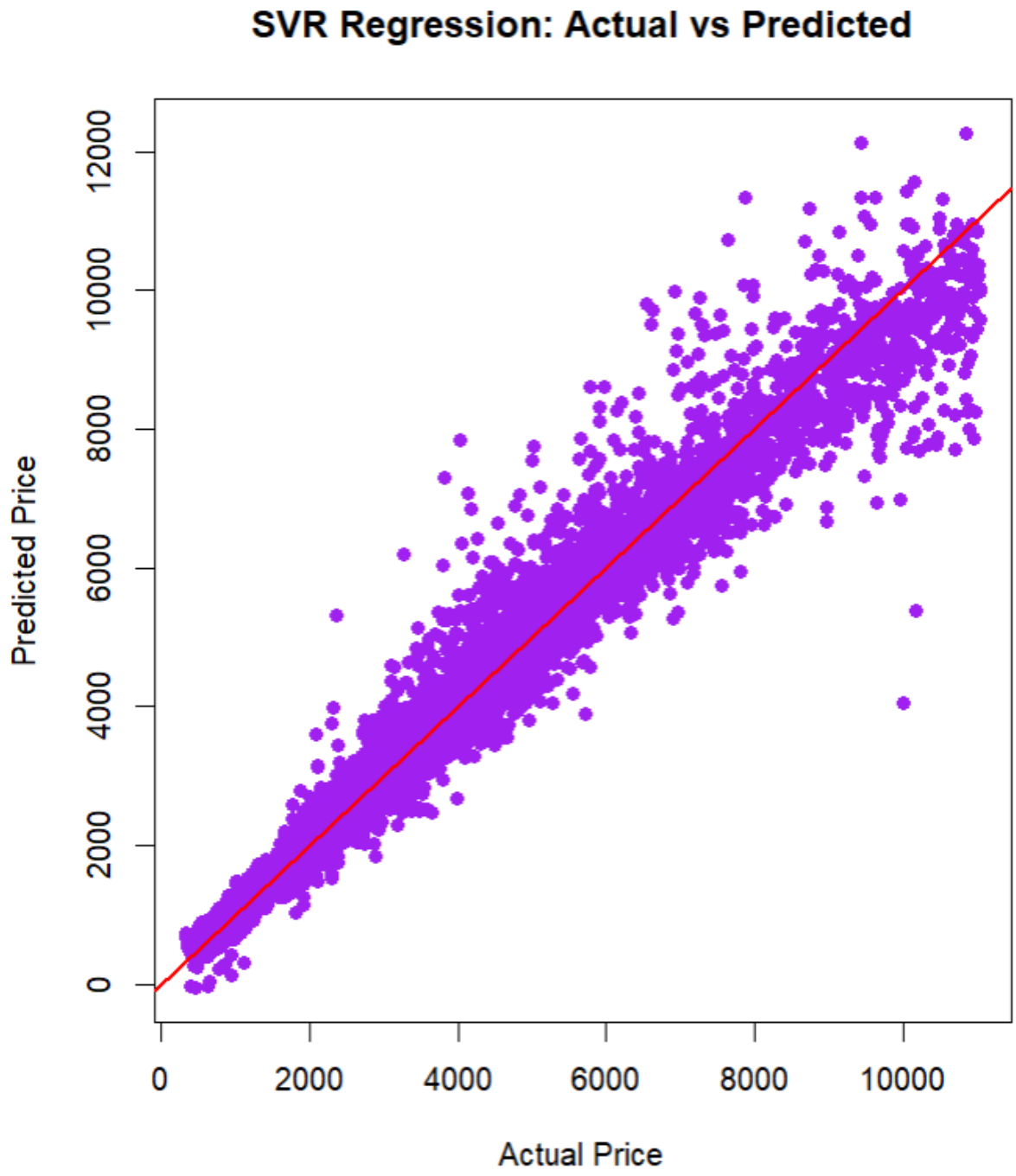
Justification for choice:

- Diamond pricing is highly non-linear, especially with respect to carat and volume
- SVR performs well on complex, high-dimensional data
- Robust against overfitting when properly configured
- Commonly used in real-world regression problems

SVR allows modeling subtle patterns that linear regression may fail to capture.

9.3.1 SVM Output

```
SVR (SVM REGRESSION) PERFORMANCE METRICS
 Accuracy (R²): 0.9721
 MAE          : 256.8297
 MSE          : 187818.7
 RMSE         : 433.3805
>
> acc_svr <- R2_svr
> acc_svr
[1] 0.9721471
```


SVR Regression: Actual vs Predicted
Predicted Price
0
2000
4000
6000
8000
10000
12000
Actual Price
0
2000
4000
6000
8000
10000

9.4 Decision Tree Regression

Decision Tree Regression was used to model price prediction through **hierarchical rule-based splits** on the input features.

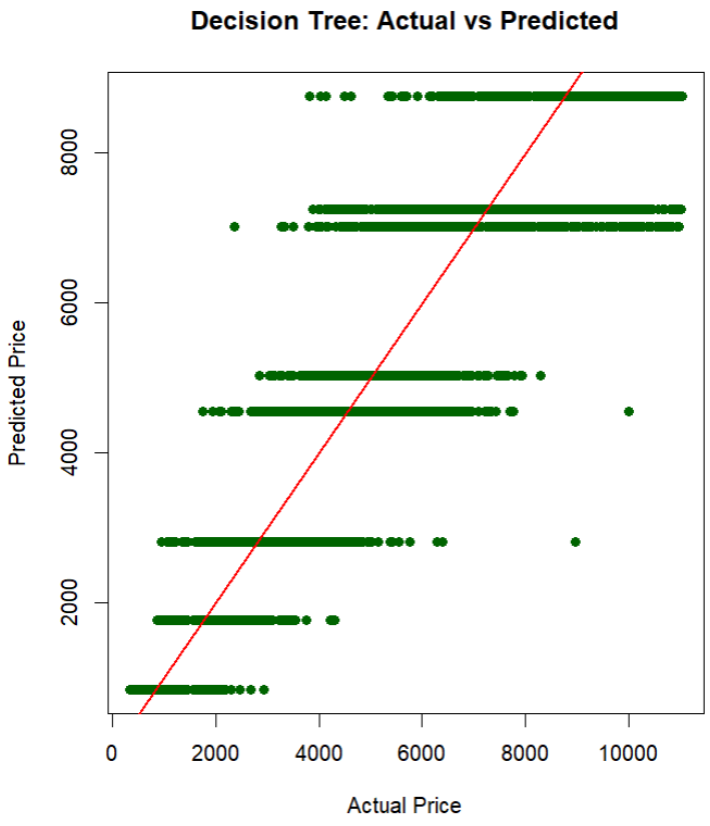
Justification for choice:

- Produces interpretable decision rules
- Handles both numerical and categorical variables naturally
- Captures non-linear interactions between features
- Useful for explaining pricing logic in business contexts

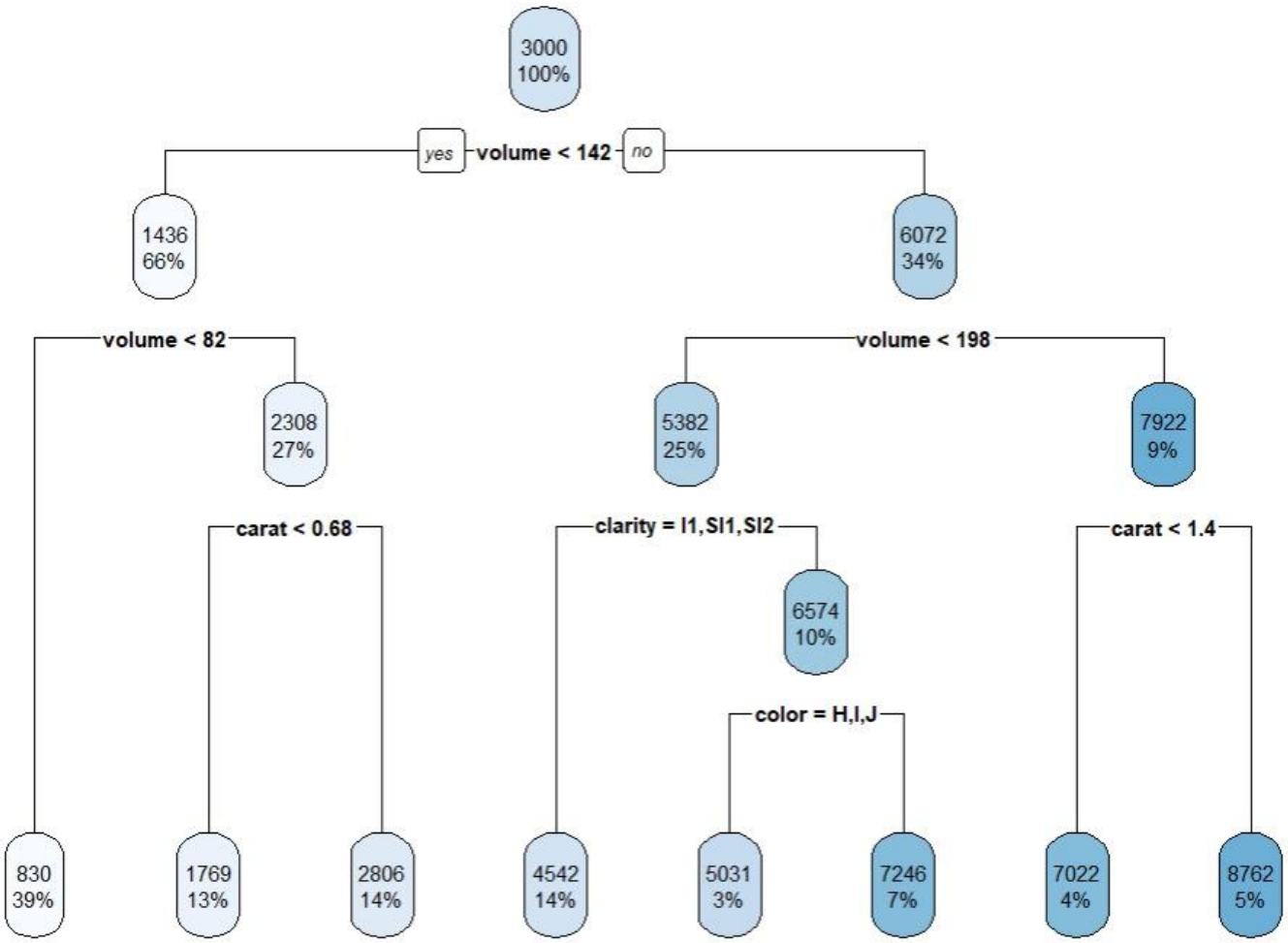
Decision trees help visualize how different diamond attributes influence pricing decisions.

9.4.1 Decision Tree Output

```
DECISION TREE REGRESSION PERFORMANCE METRICS
 Accuracy (R²): 0.9001
 MAE          : 533.4934
 MSE          : 673160.7
 RMSE         : 820.4637
>
> # Extract accuracy alone
> acc_tree <- R2_tree
> acc_tree
[1] 0.9000915
```



Decision Tree for Diamond Price



10.0 DISCUSSION/QUANTIFICATION OF PROJECT FINDINGS

This section discusses the key findings obtained from data preprocessing, exploratory data analysis, hypothesis testing, and predictive modeling. Quantitative results are used to support conclusions and justify modeling decisions.

10.1 Key Insights from Data Analysis

The exploratory data analysis and hypothesis testing revealed that **diamond price is primarily driven by size-related attributes** rather than proportional measurements or single quality metrics.

- **Carat** showed a very strong positive correlation with price ($r \approx 0.925$), confirming it as the most influential feature.
- The engineered feature **volume** demonstrated a similarly strong correlation ($r \approx 0.926$), validating its inclusion as a predictor.

10.2 Model Performance Comparison (Quantitative Evaluation)

Three regression-based models were trained and evaluated using identical training and testing datasets.

Model	**R^2 (Accuracy)**	**RMSE (Approx.)**	**Interpretation**
Linear Regression	~0.92	~718	Strong baseline, linear assumption
Decision Tree Regression	~0.90	~820	Interpretable but less precise
Support Vector Regression (SVR)	**~0.97**	**~433**	Best overall performance

Key observations:

- **SVR achieved the highest accuracy**, explaining approximately **97% of the variance** in diamond prices.

- SVR reduced prediction error by **~40% compared to linear regression** and **~47% compared to decision trees** (based on RMSE).
- Decision trees produced step-wise predictions, limiting accuracy despite interpretability.

10.3 Interpretation of Model Behavior

- **Linear Regression** effectively captured global trends but failed to model non-linear price growth at higher carat values.
- **Decision Tree Regression** provided clear pricing rules but suffered from over-simplification, predicting fixed price levels.
- **Support Vector Regression** successfully modeled non-linear relationships between price and diamond characteristics, leading to superior performance.

Visual inspection of **Actual vs Predicted plots** further confirmed that SVR predictions clustered most closely around the ideal diagonal line.

10.4 Practical Implications

The findings suggest that:

- Diamond pricing systems should prioritize **size-based metrics (carat, volume)**.
- Quality attributes (cut, color, clarity) should be used as **adjustment factors** rather than primary drivers.
- Advanced non-linear models such as **SVR** are more suitable for real-world diamond price prediction than simple linear or rule-based models.

11.0 OVERALL PROJECT CONCLUSION

This project successfully applied the Big Data Analytics workflow to a real-world dataset. Through systematic preprocessing, statistical analysis, and model comparison, meaningful insights were extracted and quantified.

The **Support Vector Regression model** was selected as the final predictive model due to its superior accuracy and robustness. The results demonstrate how combining statistical reasoning with machine learning techniques leads to reliable and interpretable predictive solutions.