Optimizing Diamond Valuation: A Machine Learning Approach to Price Prediction

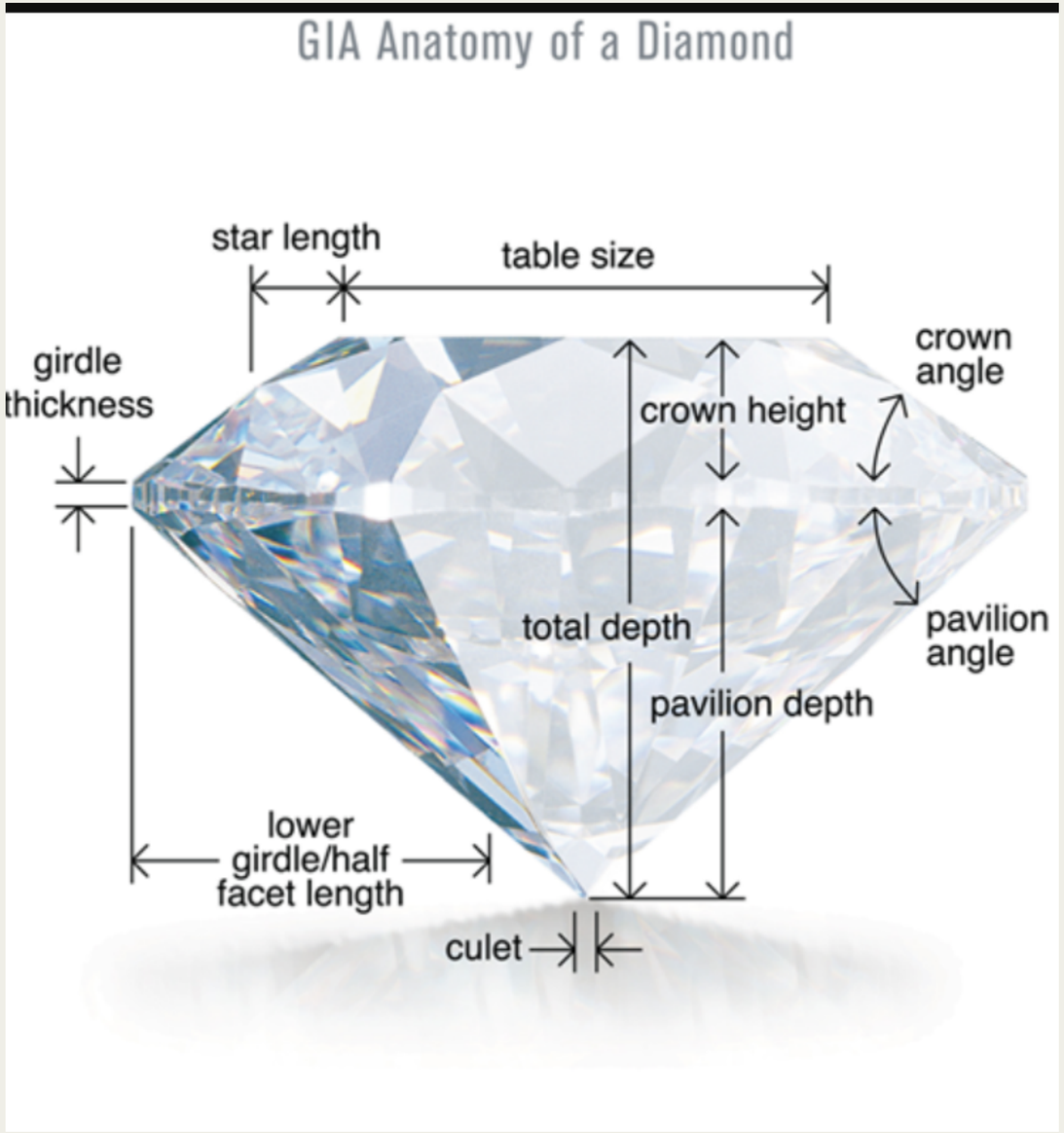
B565: DATA MINING

Project Members:

- Jui Ambikar
- Kasturi Disale
- Venkata Hari Chandan Vemuganti

AGENDA

- Domain Overview
- Motivation & Goal
- Dataset Overview
- Literature Study
- Exploratory Data Analysis
- Feature Engineering
- Model Comparison
- Model Tuning
- Future Scope


GIA Anatomy of a Diamond
star length
table size
girdle thickness
crown angle
crown height
total depth
pavilion angle
pavilion depth
lower girdle/half facet length
culet

Domain Knowledge

Traditional diamond valuation relies on subjective assessments based on the "4 Cs" – cut, color, clarity, and carat weight. Cut assesses the diamond's shaping and faceting, impacting its brilliance. Color grades range from colorless to light yellow; less color increases value. Clarity considers internal and external flaws, while carat weight reflects size. Expert gemologists apply subjective judgments, leading to pricing variations. The introduction of a machine learning approach aims to bring objectivity and precision by analyzing a comprehensive dataset.

MOTIVATION & GOAL

Traditional diamond valuation methods, relying on subjective assessments of the "4 Cs," have limitations due to variations in individual interpretations and subjective judgments by gemologists. This subjectivity can lead to inconsistencies in pricing and may not fully capture the complex factors influencing diamond values. The motivation for introducing a machine learning approach is to bring more objectivity, consistency, and precision to the valuation process.

The dataset utilized is recent, representing a new collection of information on diamond prices, and there is very limited prior analysis or research conducted on it, providing an opportunity for novel insights and contributions.

The project aims to improve diamond valuation accuracy through a machine learning regressor model. The ultimate goal is to contribute to quantitative gemology and transform the diamond valuation process.

DATASET OVERWIEW

- This dataset has 219703 graded diamonds with 25 columns of charecteristic information.

- The columns present are cut, color, clarity, carat_weight, cut_quality, lab, symmetry, polish, eye_clean, culet_size, culet_condition, depth_percent, table_percent, meas_length, meas_width, meas_depth, girdle_min, girdle_max, fluor_color, fluor_intensity, fancy_color_dominant_color, fancy_color_secondary_color, fancy_color_overtone, fancy_color_intensity, total_sales_price

- The data is sourced from Kaggle.com.

LITERATURE REVIEW

- Alsuraihi et al. (2020): Utilization of Random Forest Regression, noted for low MAE and RMSE but lacked consideration of class imbalance and diamond cut’s impact.

- Mihir et al. (2021): Emphasized CatBoost Regression's effectiveness, suggesting additional attributes for accuracy.

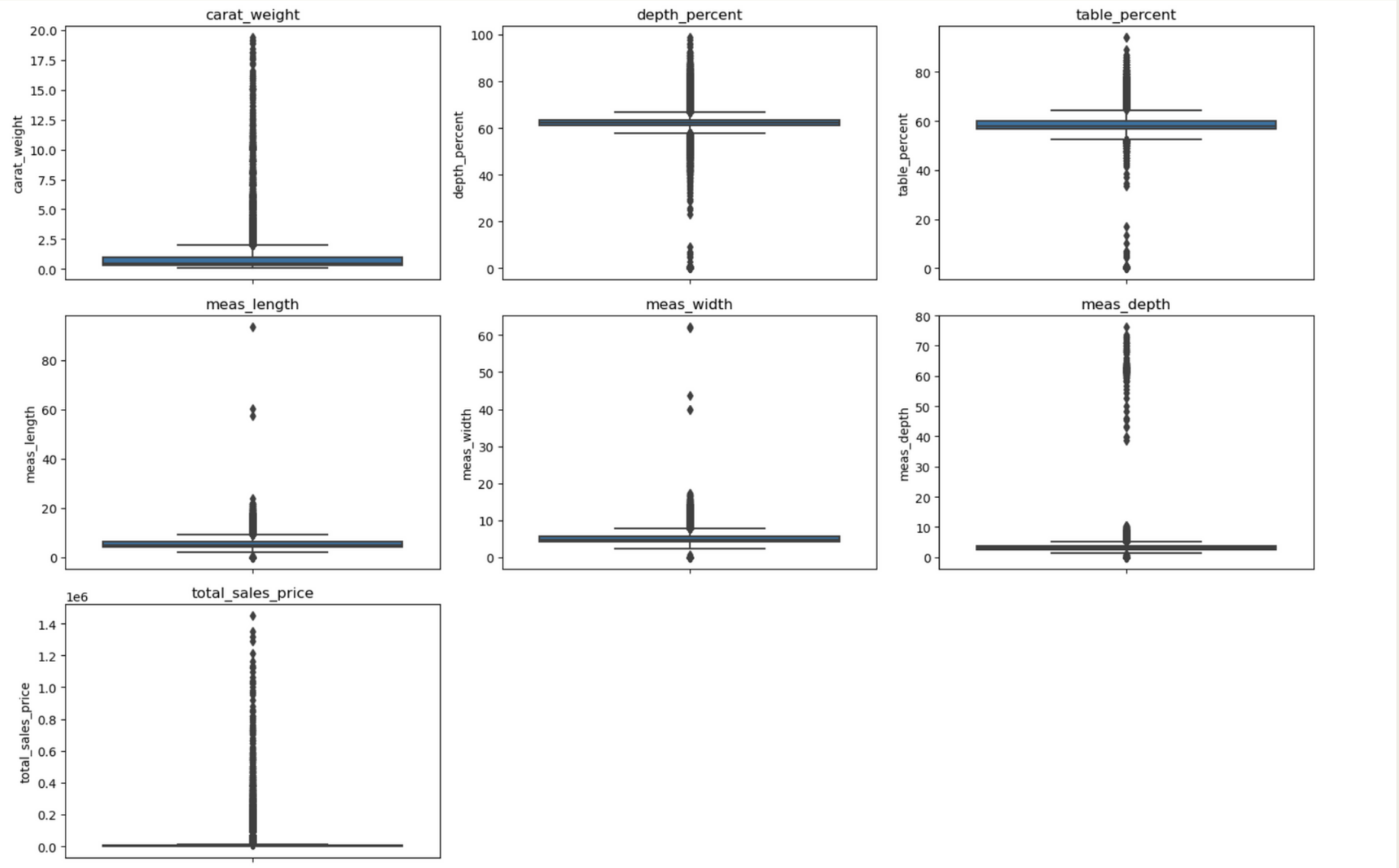
- Kigo et al. (2023): Comprehensive use of supervised learning algorithms, with Random Forest showing high accuracy, indicated by an R2 score of 0.985.

- Mamonov and Triantoro (2022): Explored e-commerce aspects using Decision Forest and ANN, highlighting primary price determinants but not using methods like XGBoost.

LITERATURE REVIEW

- Common Trends: Preference for ensemble and advanced regression techniques.
- Identified Gaps: Need for diverse feature inclusion, addressing dataset imbalances, and employing varied evaluation metrics.
- Future Directions: Exploring robust methods like XGBoost, incorporating novel machine learning or deep learning algorithms, and thorough evaluation using multiple regression metrics.

EDA: OUTLIER ANALYSIS



EDA: OUTLIER ANALYSIS

- **Carat Weight:** There are visible outliers, with some diamonds having significantly higher carat weights than the majority.
- **Depth Percent:** This column also shows outliers, particularly on the higher end.
- **Table Percent:** There are outliers present on both lower and higher ends.
- **Measurements (Length, Width, Depth):** Each of these columns has outliers, especially on the higher end, indicating some diamonds are much larger in size than most.
- **Total Sales Price:** There are significant outliers, with some diamonds having exceptionally high sales prices.

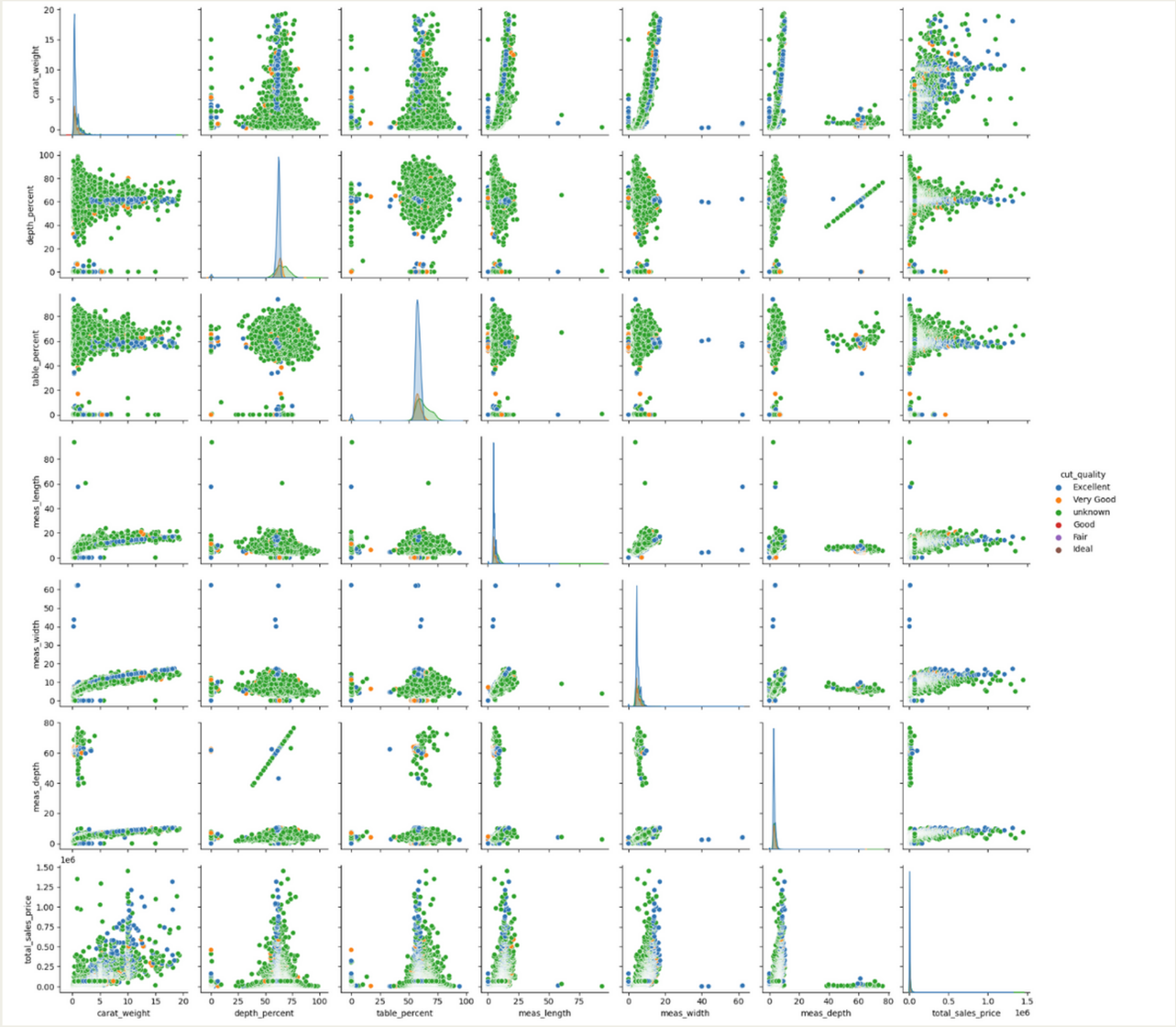
EDA: OUTLIER ANALYSIS

The ***IqR method***, which is based on quartiles, tends to be less sensitive than the MAD method, which is evident from the generally lower counts of outliers.

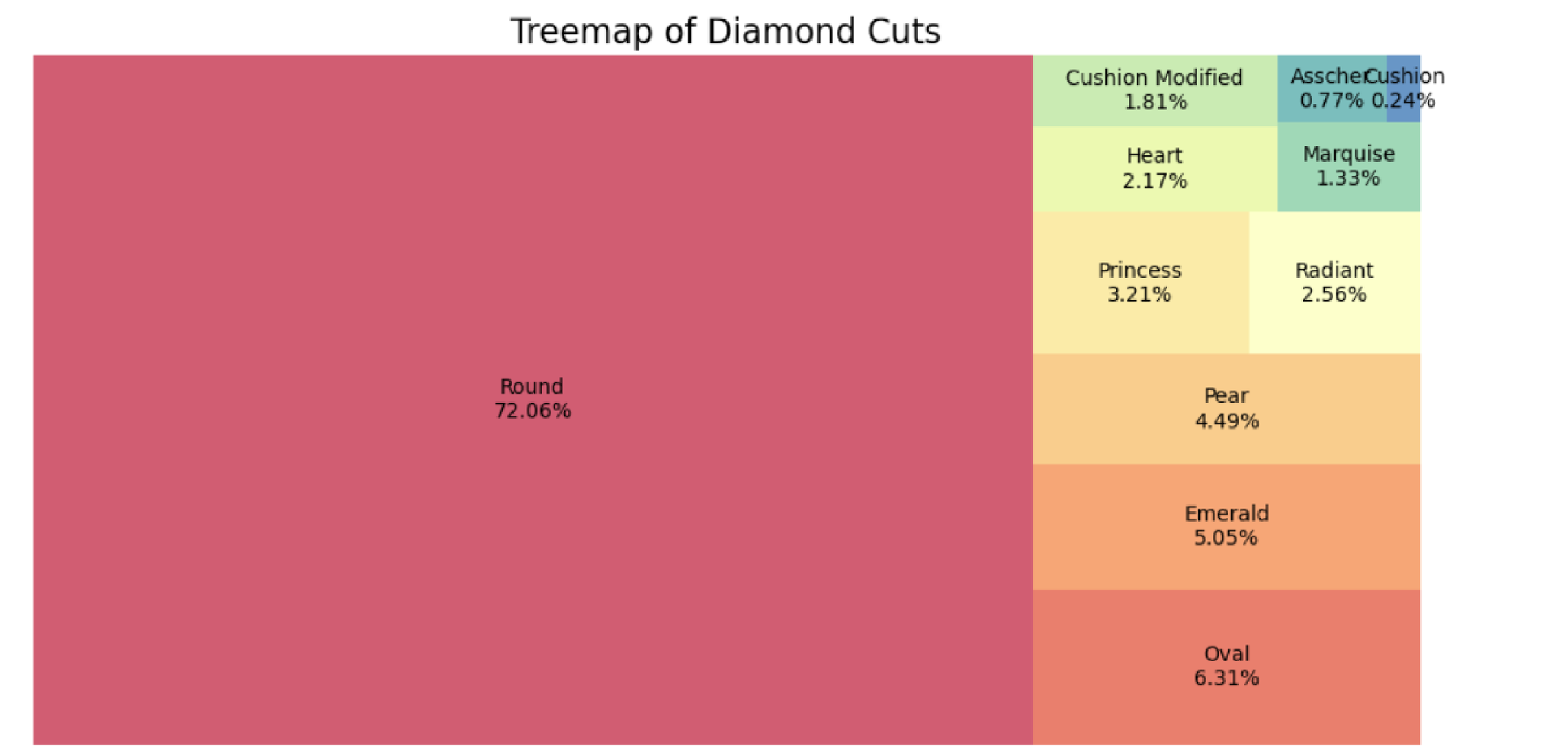
The ***MAD (Median Absolute Deviation) method***, which is less influenced by extreme values, identifies a larger number of outliers in most columns. This could be due to the presence of extreme values or heavy tails in the data distribution. Depth Percent, Table Percent, and Total Sales Price show a high number of outliers in both methods, indicating these attributes might have a wide range of values or distributions that deviate from normality.

Identified **domain-specific** outliers based on certain industry thresholds: Carat Weight: Outliers were considered for weights above 2.5 carats, as such large diamonds are relatively rare. Depth Percent: Typical depth percent ranges between 55% and 70%. Values outside this range were considered outliers. Table Percent: A typical range is between 53% and 65%. Values outside this were also flagged as outliers.

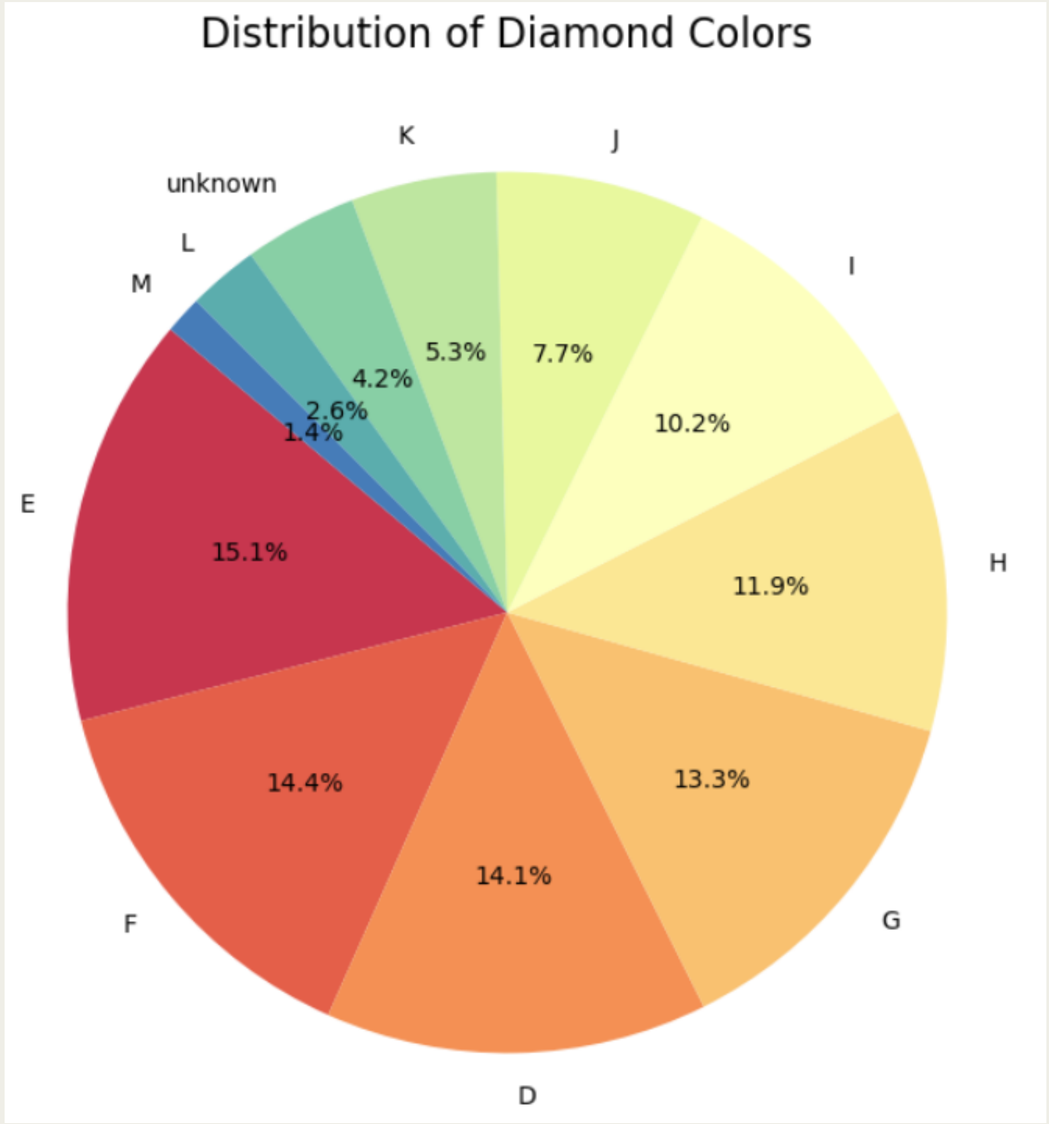
EDA: PAIR PLOT



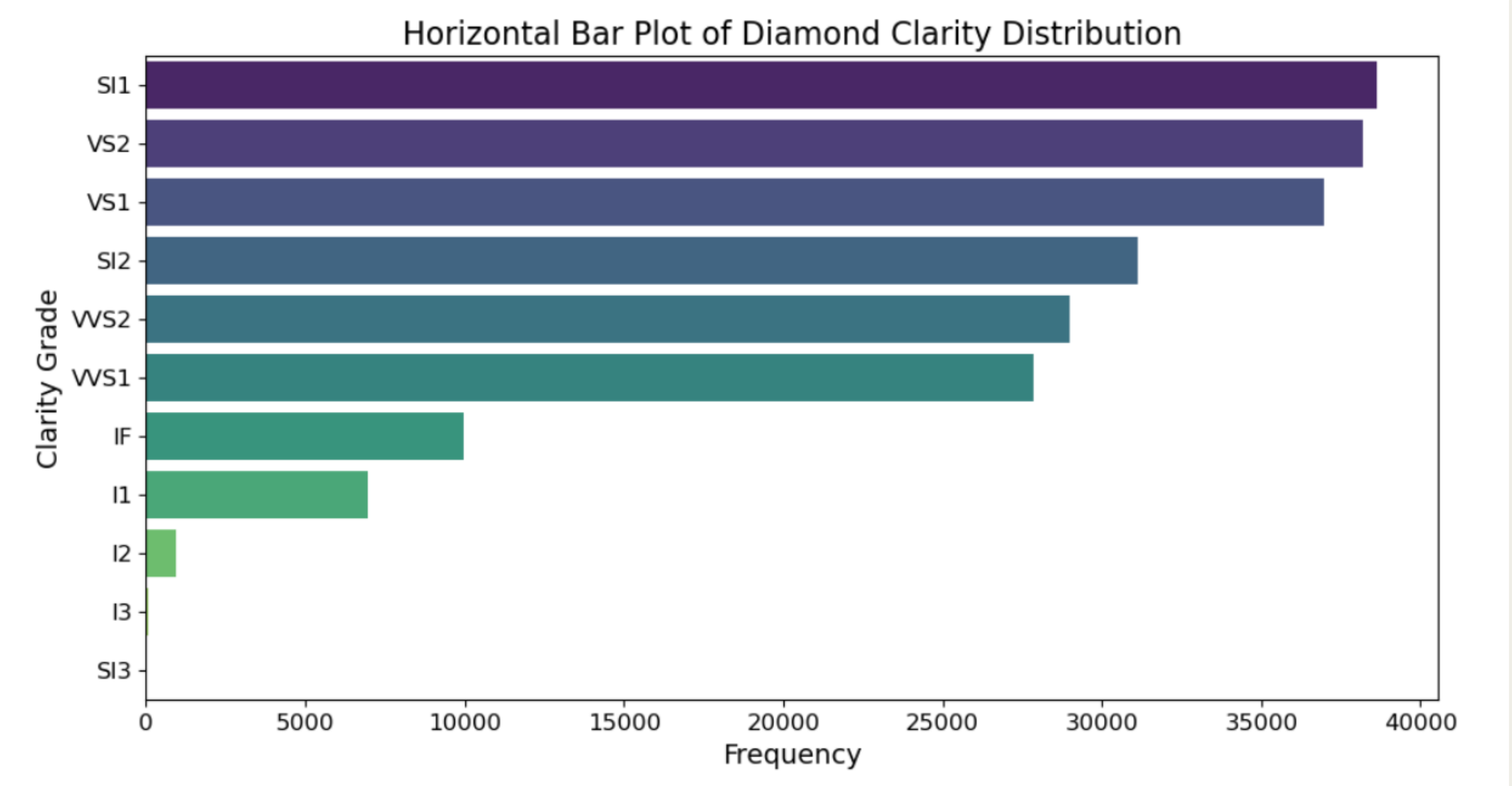
EDA: 4C'S ANALYSIS



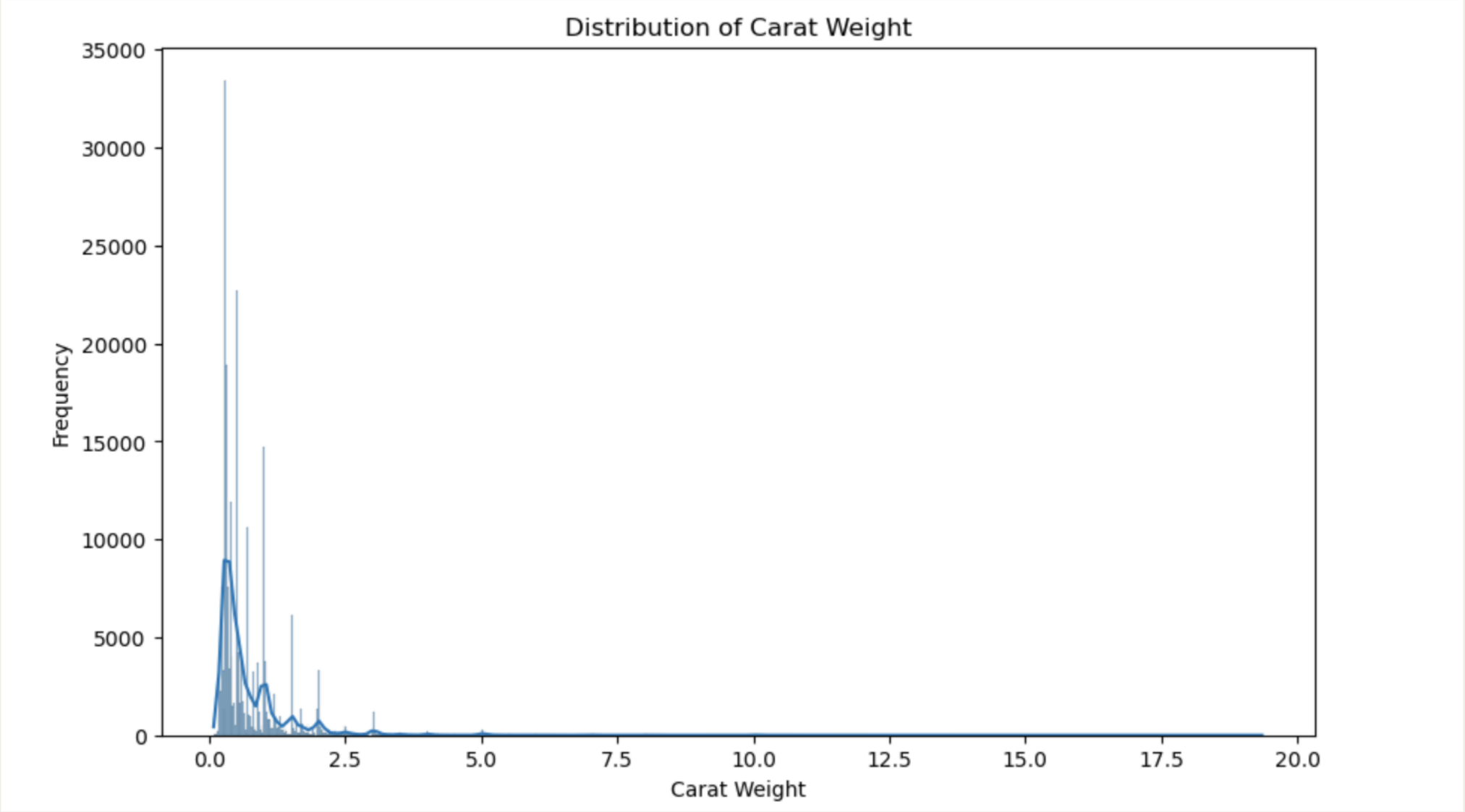
EDA: 4C'S ANALYSIS



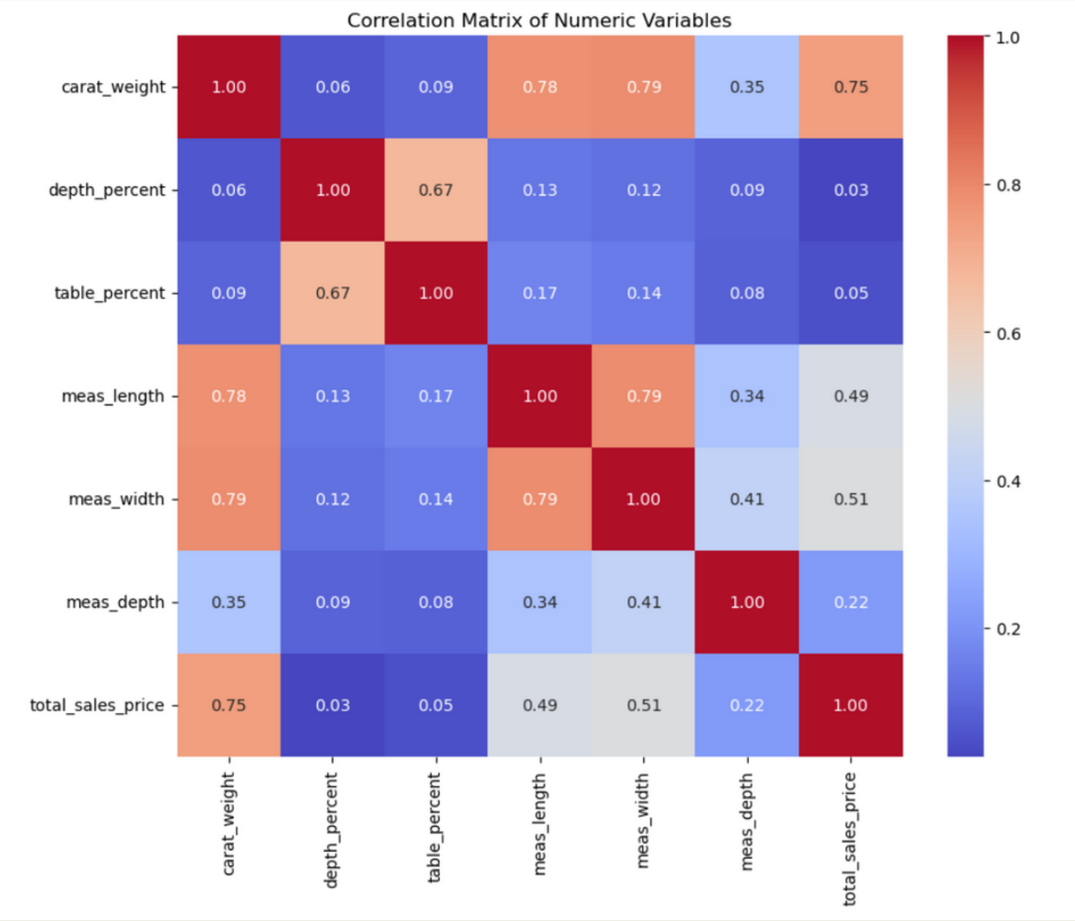
EDA: 4C'S ANALYSIS



EDA: 4C'S ANALYSIS



EDA: CORRELATION PLOT



SCALING: ROBUST SCALER

1. Outliers in Data: The dataset has significant outliers, especially in key features like 'carat_weight' and 'total_sales_price', which can distort the performance of many algorithms.
2. Disparate Feature Scales: Numeric features have widely varying scales, potentially leading to biased results in models where larger-scale features could dominate.
3. Robustness of Robust Scaler: Specifically designed to handle outliers, the Robust Scaler uses median and interquartile range for scaling, reducing the influence of extreme values.
4. Normalizing Feature Contributions: Scaling ensures each feature contributes equally to the analysis, crucial for distance-based and assumption-dependent algorithms.
5. Maintaining Data Structure: Unlike some transformations, scaling with the Robust Scaler preserves the underlying distribution of data, maintaining the intrinsic properties of the dataset.

SCALING: ONE HOT ENCODING

1. Nominal Categories: Ideal for features like 'cut', 'color', and 'clarity', which are nominal with no intrinsic order, ensuring equal treatment of all categories.
2. Model Compatibility: Transforms categorical data into a numerical format, making it suitable for various machine learning algorithms that require numerical input.
3. Information Preservation: Maintains the distinct value of each category, avoiding any loss of information, unlike ordinal encoding methods.
4. Prevents Misinterpretation: Avoids misleading the model into assuming natural ordering between categories, crucial for features like 'color' and 'cut_quality'.
5. Standard Approach: A widely used and understood method, one-hot encoding helps in handling categorical variables effectively, though it may increase dataset dimensionality.

MODEL COMPARISON

The comparison of different machine learning models reveals promising performance metrics:

1. **Baseline Linear Regression (R2: 0.642, RMSE: 16669.82)**
2. **Random Forest (R^2: 0.80, MSE: 152,466,982):**
    - Demonstrates substantial accuracy in predicting diamond values.
3. **XGBoost (R^2: 0.87, MSE: 102,589,108):**
    - Outperforms other models, indicating superior predictive capabilities.
4. **CatBoost (R^2: 0.86, MSE: 111,041,062):**
    - Achieves high accuracy, contributing to the overall success of the project.
5. **LightGBM (R^2: 0.81, MSE: 148,254,912):**
    - Exhibits commendable performance in predicting diamond prices.
6. **SVM (R^2: 0.04, MSE: 743,253,827):**
    - While SVM shows a lower R^2, it provides valuable insights into the complexities of the dataset.

MODEL TUNING

Hyperparameter-tuned Xgboost regressor:

- Once the GridSearchCV has been applied to the training data, the resulting output displays the hyperparameters that are best suited for enhancing the model's performance.
- The optimal setup comprises of 'colsample_bytree': 0.7, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 300, and 'subsample': 0.9.
- The corresponding R^2 score, which serves as a measure of the model's accuracy, is 0.8898, indicating that the diamond pricing model is highly accurate.

Hyperparameter-tuned Catboost regressor:

- After performing GridSearchCV on the training data, the results suggest the best hyperparameters to optimize the performance of the CatBoost Regressor.
- The optimal configuration includes 'border_count': 128, 'depth': 8, 'iterations': 100, 'l2_leaf_reg': 1, and 'learning_rate': 0.1.
- The corresponding R^2 score, which indicates the model accuracy, is 0.8622. This suggests that the model is highly accurate in predicting diamond pricing.

CONCLUSION

The project has successfully implemented a machine learning-based regression model to enhance diamond valuation accuracy. The results showcase the potential of data-driven methodologies in overcoming the limitations of traditional approaches.

FUTURE SCOPE

- **What-If Analysis Integration:** Integrate advanced "what-if analysis" functionalities to the diamond pricing prediction model. This would allow users to simulate various scenarios and understand the impact of changing parameters on diamond prices.
- **Enhanced Feature Engineering:** Continuously explore and incorporate additional features into the model. This could include factors such as geopolitical events, global economic indicators, and emerging trends in diamond preferences.
- **Website and User Interface Development:** Create a user-friendly website or a dedicated platform where users, such as diamond traders, retailers, or consumers, can easily access the prediction tool. The interface should provide intuitive visualizations, historical pricing trends, and interactive tools for users to input parameters and obtain predicted prices.
- **Expand to Other Gemstones and Precious Metals:** Expanding the scope of the project to include the prediction of prices for other gemstones and precious metals.


Thank you!