Diamond price prediction 2024 Project

SHAI Training 2024

Team Names:

Mays Moh’d Al-Fasfous

Hesham Saad Alsaadi

Fahed Shadid

Sarah Hasan

1. Introduction

In the realm of data analytics, the predictive modeling of diamond prices stands as a quintessential exercise, offering a rich tapestry of data attributes to explore and analyze. This document encapsulates a comprehensive journey through the realms of data preprocessing, exploratory data analysis (EDA), model building, and submission preparation within the context of predicting diamond prices.

The dataset under scrutiny encompasses a diverse array of features, ranging from intrinsic characteristics such as carat weight, cut quality, color, and clarity to physical dimensions like length, width, and depth. With nearly 54,000 diamonds at our disposal, this dataset presents an invaluable opportunity for budding data scientists and seasoned analysts alike to delve deep into the nuances of predictive modeling.

This document serves as a roadmap, guiding the reader through each phase of the analytical process, from initial data exploration to the fine-tuning of predictive models. Along the way, we'll uncover insights, address challenges, and ultimately strive to construct a robust model capable of accurately predicting diamond prices.

1.1 Background

The diamond industry has long captivated both consumers and investors with its allure, representing not only a symbol of enduring love but also a lucrative market for investors and traders. Understanding the factors influencing diamond prices is of paramount importance in this industry, as it enables stakeholders to make informed decisions regarding pricing, marketing, and investment.

In this backdrop, the dataset containing prices and attributes of over 54,000 diamonds emerges as a treasure trove for data analysts and enthusiasts. By leveraging advanced analytics techniques, we aim to unravel the intricate relationships between diamond attributes and prices, providing actionable insights for industry stakeholders.

1.2 Objectives

The primary objectives of this document are as follows:

1- Explore the dataset: Conduct exploratory data analysis (EDA) to gain insights into the distribution, correlation, and significance of various diamond attributes.
2- Preprocess the data: Cleanse, transform, and engineer features to enhance the quality and relevance of input data for modeling.
3- Build predictive models: Utilize machine learning algorithms to develop predictive models capable of accurately estimating diamond prices.
4- Evaluate model performance: Assess the performance of the developed models using appropriate evaluation metrics and techniques.
5- Generate submissions: Prepare submissions in the required format for the competition, incorporating the insights gained and models developed throughout the analysis.

2- Dataset Overview

Our project revolves around a dataset comprising data on over 54,000 diamonds. It includes essential attributes such as carat weight, cut quality, color, clarity, and physical dimensions alongside their corresponding prices. This real-world dataset serves as the foundation for our data analytics efforts, aiming to extract insights and build predictive models for accurate diamond price estimation.

3- Framing the Problem

In the realm of diamond pricing, understanding the intricate interplay of various factors is paramount. Leveraging domain knowledge reveals that diamond prices are influenced by several key attributes, often encapsulated by the famous "4Cs" - Cut, Color, Clarity, and Carat Weight. Each of these factors contributes uniquely to the overall value and desirability of a diamond.

Utilizing this domain knowledge, the task at hand is to develop a predictive model capable of accurately estimating diamond prices based on these fundamental attributes. By framing the problem in this context, we aim to explore the relationships between the 4Cs and diamond prices, discerning patterns that can guide pricing strategies and market analysis.

Through data analytics techniques, we seek to uncover insights that elucidate the nuances of diamond pricing dynamics, ultimately empowering stakeholders to make informed decisions in the vibrant diamond market.

4- Selecting a Performance Measure

To gauge the effectiveness of our predictive models in estimating diamond prices, we opt for Root Mean Squared Error (RMSE) as our performance measure. RMSE calculates the average disparity between predicted and actual prices. By minimizing RMSE, we strive to enhance the accuracy of our models across various diamond attributes, including cut, color, clarity, and carat weight. This choice ensures our models meet the stringent standards demanded by the dynamic diamond market.

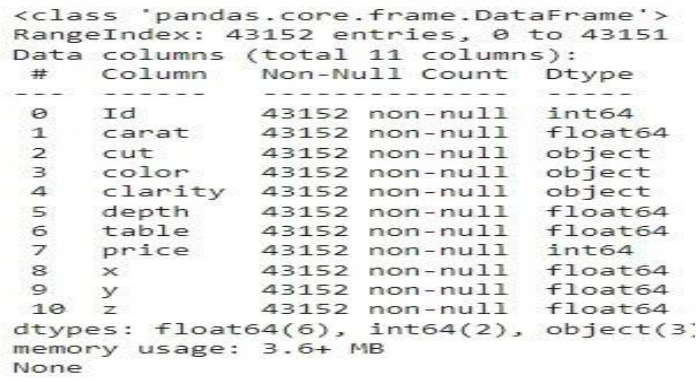
5- Exploratory Data Analysis (EDA)

Exploring the dataset is a fundamental step in understanding its characteristics and preparing for further analysis. Here's an overview of our exploratory analysis:

5.1 Previewing Data: We start by examining the first few rows of the dataset to get a glimpse of its structure and content.

5.2 Inspecting Tail End: Similarly, we review the last few rows to ensure data consistency and observe any potential patterns.

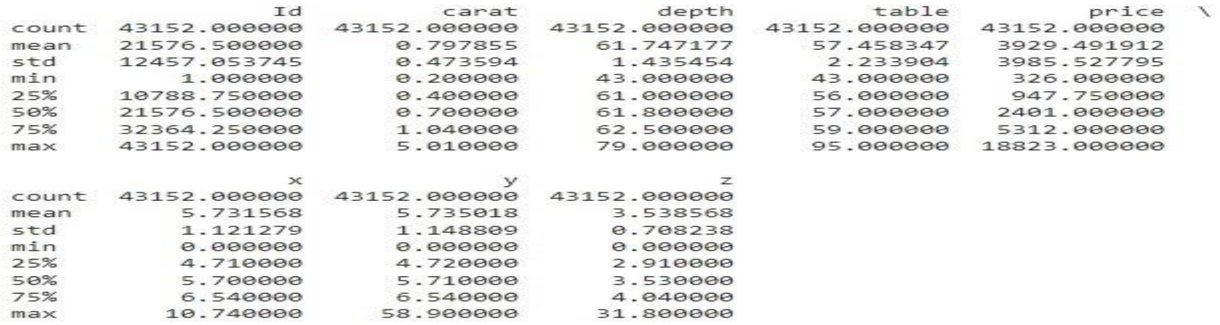
5.3 Data Summary: Utilizing the info() function, we obtain a summary of the dataset, including the data types and non-null counts for each column as in below:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43152 entries, 0 to 43151
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Id        43152 non-null  int64
 1   carat     43152 non-null  float64
 2   cut       43152 non-null  object
 3   color     43152 non-null  object
 4   clarity   43152 non-null  object
 5   depth     43152 non-null  float64
 6   table     43152 non-null  float64
 7   price     43152 non-null  int64
 8   x         43152 non-null  float64
 9   y         43152 non-null  float64
 10  z         43152 non-null  float64
dtypes: float64(6), int64(2), object(3)
memory usage: 3.6+ MB
None
```

5.4 Number of Instances and Features: We ascertain the total number of instances and features in the dataset. In our dataset : Number of instances: 43152, Number of features: 11

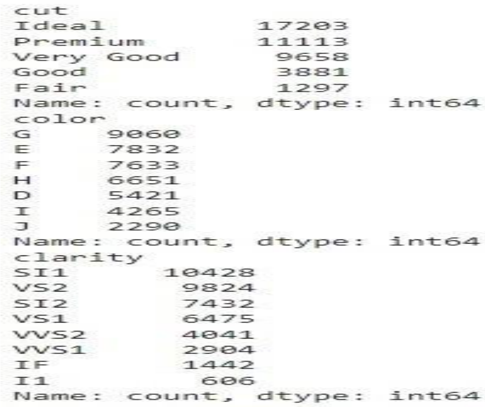
5.5 Statistical Summary: Employing describe(), we generate descriptive statistics for numerical features, revealing insights into central tendency, dispersion, and distribution as in below:



```
                 Id         carat         depth         table         price  \
count  43152.000000  43152.000000  43152.000000  43152.000000  43152.000000
mean   21576.500000      0.797855     61.747177     57.458347   3929.491912
std    12457.053745      0.473594      1.435454      2.233904   3985.527795
min        1.000000      0.200000     43.000000     43.000000    326.000000
25%    10788.750000      0.400000     61.000000     56.000000    947.750000
50%    21576.500000      0.700000     61.800000     57.000000   2401.000000
75%    32364.250000      1.040000     62.500000     59.000000   5312.000000
max    43152.000000      5.010000     79.000000     95.000000  18823.000000

                  x             y             z
count  43152.000000  43152.000000  43152.000000
mean       5.731568      5.735018      3.538568
std        1.121279      1.148809      0.708238
min        0.000000      0.000000      0.000000
25%        4.710000      4.720000      2.910000
50%        5.700000      5.710000      3.530000
75%        6.540000      6.540000      4.040000
max       10.740000     58.900000     31.800000
```

5.6 Categorical Feature Counts: We explore the frequency distribution of categorical features such as cut, color, and clarity to discern any prevalent categories.



```
cut
Ideal        17203
Premium      11113
Very Good     9658
Good          3881
Fair          1297
Name: count, dtype: int64
color
G    9060
E    7832
F    7633
H    6651
D    5421
I    4265
J    2290
Name: count, dtype: int64
clarity
SI1     10428
VS2      9824
SI2      7432
VS1      6475
VVS2     4041
VVS1     2904
IF       1442
I1        606
Name: count, dtype: int64
```

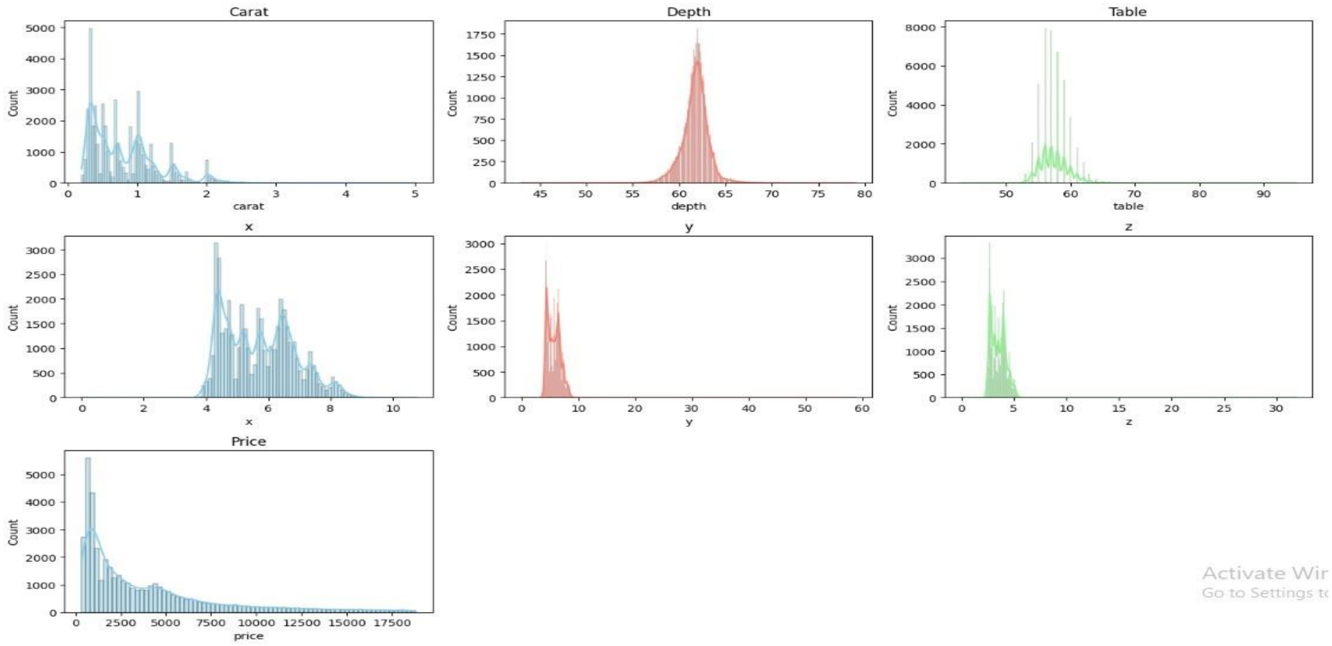
5.7 Handling Missing Values: We verify the absence of missing values or null data within the dataset. Our dataset does not contain any missing values or null data.

5.8 Duplicate Entries Check: Lastly, we ensure data integrity by confirming the absence of duplicate entries. Our dataset is free from duplicate entries, ensuring that each record is unique and contributes distinct information for analysis and modeling.

6- Data Analysis and Visualization

6.1 Visualize distributions of numerical features

To gain further insights into the distribution of numerical features in the diamonds dataset, we visualize these distributions using histograms. This step helps us understand the underlying patterns, detect outliers, and identify skewness in the data.



Observations from the Distributions:

1. Carat:
    - X-axis: The carat values start from just above 0 and extend to just before 3. There are outliers extending to values greater than 5.
    - Y-axis: The frequency of carat values ranges from 0 to approximately 5000.
    - Observation: The majority of diamonds have a carat value below 3, with a few outliers exceeding 5.
2. Depth:

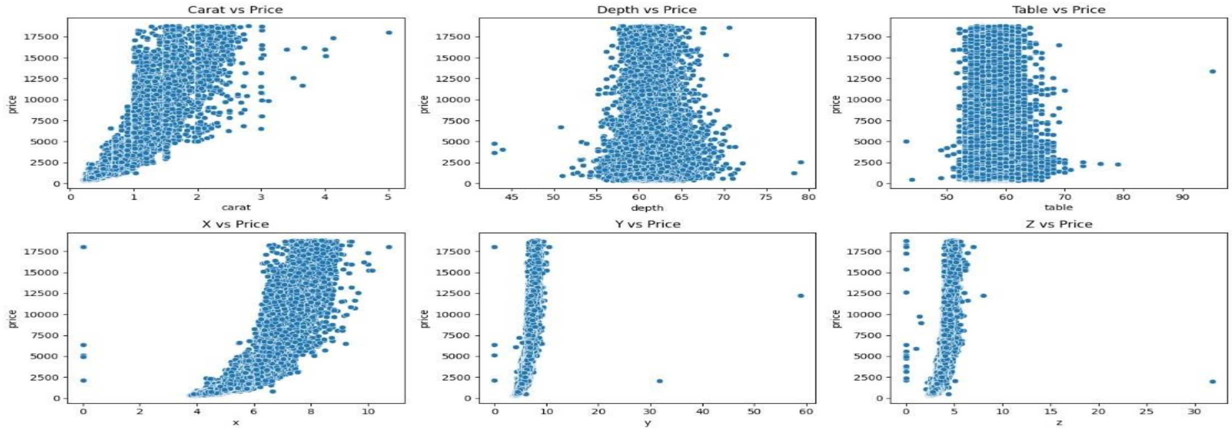
- X-axis: Depth values range from 55 to just before 70, with outliers extending up to 80.
- Y-axis: The frequency ranges from 0 to 1750.
- Observation: Most diamonds have a depth percentage between 55 and 70, with a few outliers reaching 80.

3. Table:
    - X-axis: Table values start from 50 and go up to just before 70, with outliers extending beyond 90.
    - Y-axis: The frequency ranges from 0 to 8000.
    - Observation: The table width for most diamonds falls between 50 and 70, with some outliers exceeding 90.

4. X (Length):
    - X-axis: Length values start from above 3 and go up to 10, with outliers extending beyond 10.
    - Y-axis: The frequency ranges from 0 to 3000.
    - Observation: Most diamonds have a length between 3 and 10 mm, with a few outliers above 10 mm.
5. Y (Width):
    - X-axis: Width values start from above 0 and go up to 10, with outliers extending to 60.
    - Y-axis: The frequency ranges from 0 to 3000.
    - Observation: The majority of diamonds have a width between 0 and 10 mm, with some outliers reaching up to 60 mm.
6. Z (Depth):
    - X-axis: Depth values start from above 0 and go up to just beyond 5, with outliers extending to 30.
    - Y-axis: The frequency ranges from 0 to just over 3000.
    - Observation: Most diamonds have a depth between 0 and 5 mm, with a few outliers reaching up to 30 mm.

7. Price:
    - X-axis: Price values start from just above 0 and go up to 17500.
    - Y-axis: The frequency ranges from 0 to just over 5000.
    - Observation: Diamond prices are concentrated between 0 and 17500 USD, with the highest frequency of diamonds priced below 17500 USD.

6.2 Visualize relationships between numerical features and target variable



Observations:

1. Carat vs. Price

Observation: There is a clear positive correlation between carat weight and price. As the carat weight increases, the price generally increases as well.

2. Depth vs. Price

Observation: The depth of the diamond does not show a strong correlation with the price. The data points are widely scattered without a clear pattern.

3. Table vs. Price

Observation: Similar to depth, the table feature shows no strong correlation with price. The scatterplot does not reveal a clear trend.

4. X (Length) vs. Price

Observation: There is a noticeable positive correlation between the length of the diamond and its price, although not as strong as the carat.

5. Y (Width) vs. Price

Observation: The width (y) of the diamond has a slight positive correlation with price, but the relationship is weaker compared to carat and length.

6. Z (Depth) vs. Price

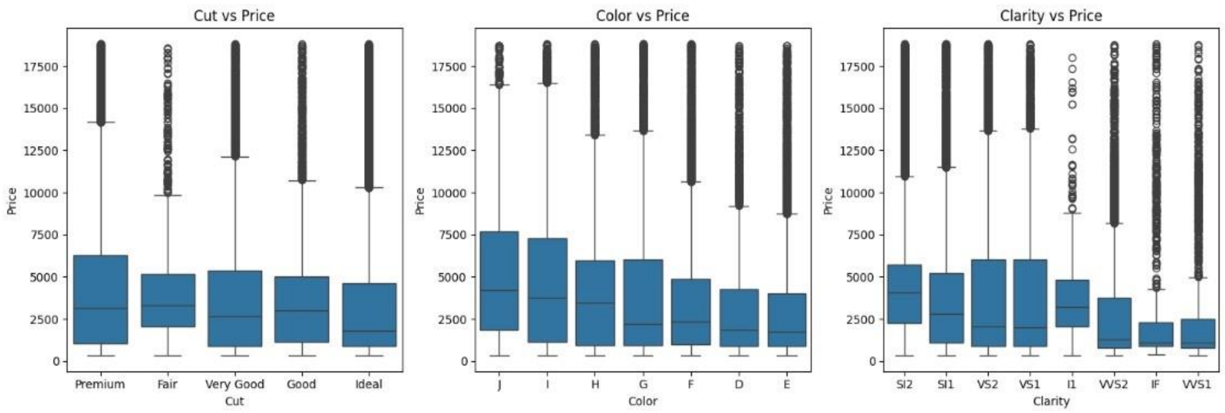
Observation: The depth (z) of the diamond shows a slight positive correlation with price, similar to the width, but with some significant outliers.

6.2.1 How These Insights Help in Model Training and Prediction:

- Feature Importance:
  - **Carat and X (Length)** are clearly the most important numerical predictors for price. These features should be given significant weight in the model.
  - **Depth, Table, Y (Width), and Z (Depth)** show less direct impact and may need to be combined with other features to extract meaningful patterns.
- The scatterplots reveal that while some features like carat and length (X) have a strong relationship with diamond price, others like depth and table do not. These insights guide us in prioritizing features, engineering new ones, and selecting appropriate models to improve the accuracy of price predictions.

6.3 Visualize relationships between categorical features and target variable

6.3.1 Box Plots



Observations:

Analyzing the mean and median prices across different cut, color, and clarity grades provides valuable insights into pricing dynamics within the diamond market.

Cut Grade:

The pricing analysis reveals varying mean and median prices across different cut grades. While 'Premium' and 'Fair' cuts exhibit the highest mean prices, the 'Ideal' cut surprisingly shows a lower mean price despite being the most prevalent. However, the median price for 'Ideal' cuts is notably lower than that of 'Premium' and 'Fair' cuts. This suggests that while 'Ideal' cuts may be more common, other factors such as rarity or demand influence pricing to a greater extent.

Color Grade:

Similarly, the analysis of color grades uncovers differences in pricing trends. Diamonds with color grade 'J' command the highest mean and median prices, indicating potential rarity-driven pricing dynamics or consumer preferences for lower color grades. Conversely, diamonds with color grades 'E' and 'D' exhibit lower mean and median prices despite their higher quality. This suggests that factors beyond color, such as cut or clarity, play a significant role in determining pricing.

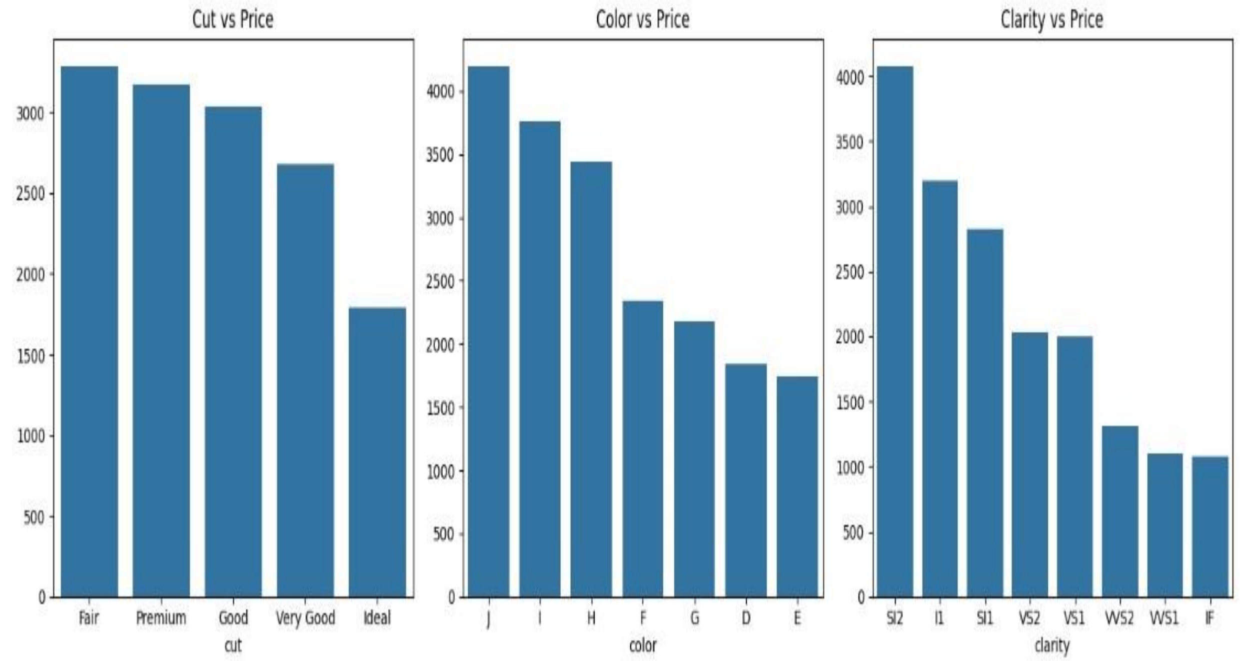
Clarity Grade:

Examining clarity grades reveals diverse pricing patterns. Diamonds with clarity grade 'SI2' exhibit the highest mean and median prices, suggesting that slight inclusions may not significantly impact pricing or consumer preferences. On the other hand, diamonds with clarity grade 'IF' (Internally Flawless) command the lowest mean and median prices, indicating that exceptional clarity may not always translate to higher pricing.

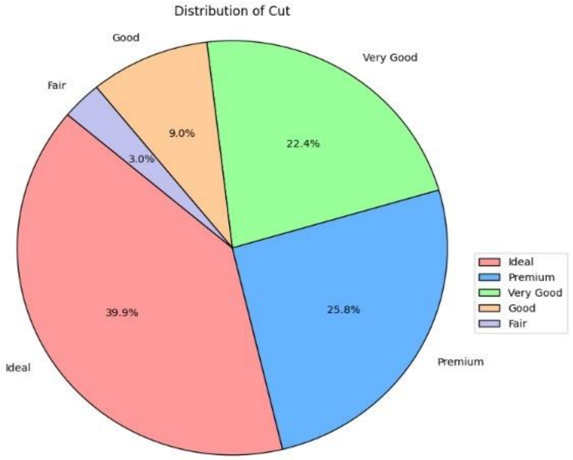
These insights highlight the complex interplay of various factors in determining diamond prices and underscore the importance of considering multiple dimensions when evaluating pricing strategies or market trends.

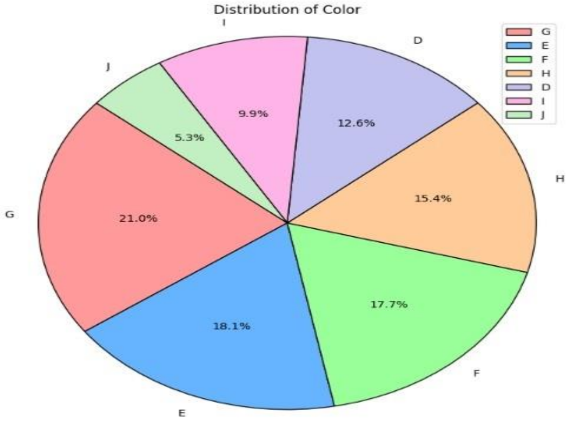
6.3.2 Visualize relationships between categorical features and target variable using bar plots and pie charts

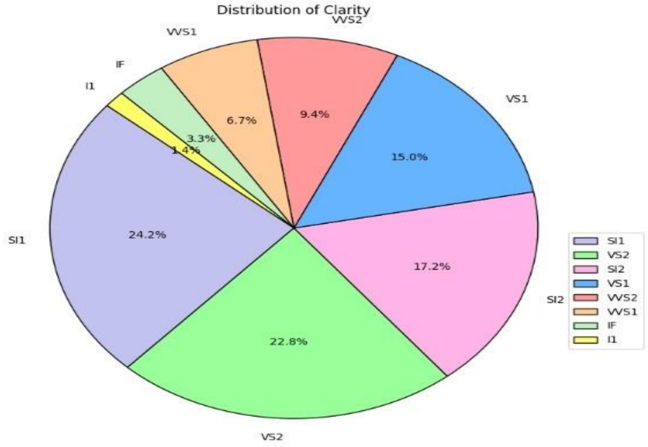
6.3.2.1 bar plots



6.3.2.2 Pie Charts







Observations:

The visualizations and statistical analyses of categorical columns provide valuable insights into pricing trends and consumer preferences in the diamond market.

Cut Distribution:

The distribution of diamond cuts highlights notable variations in pricing across different cut grades. While 'Ideal' cuts dominate the dataset, they surprisingly exhibit a lower median price compared to 'Premium' cuts. This suggests that customers may prioritize other factors over cut grade when making purchasing decisions, such as carat weight or clarity. Additionally, 'Fair' cuts, despite being the least common, show a higher median price compared to 'Good' cuts, indicating potential rarity-driven pricing dynamics.

Color Variation:

Analyzing color grades alongside pricing statistics reveals intriguing patterns in consumer preferences. Despite 'J' being the least prevalent color grade, diamonds with this grade exhibit the highest median price. Conversely, diamonds with color grades 'E' and 'D' have lower median prices despite their higher prevalence. This suggests that factors beyond color grade, such as cut quality or clarity, may significantly influence pricing decisions.

Clarity Insights:

The clarity grades exhibit a diverse pricing landscape, with notable differences in median prices across different grades. Diamonds with clarity grade 'VVS1' surprisingly show a lower median price, indicating that other factors may outweigh clarity in determining pricing. Conversely, diamonds with clarity grades 'SI2' and 'SI1' command higher median prices, suggesting that consumers may prioritize clarity over other characteristics.

Overall, these insights highlight the complex interplay between diamond features and pricing, suggesting that a combination of factors influences customer preferences and willingness to pay. As a data analyst, it's essential to consider these nuances when developing pricing strategies or analyzing market trends in the diamond industry.

6.4 Descriptive Analytic Techniques

6.4.1 Measures of Position for 'x', 'y' and 'z' columns

In our analysis of the diamonds dataset, we applied various descriptive analytic techniques to summarize and understand the data. Here, we present the measures of position for the 'x', 'y', and 'z' columns, which represent the physical dimensions of the diamonds. These measures include the 25th percentile (Q1), the 75th percentile (Q3), the Interquartile Range (IQR), and the minimum and maximum values.

These measures provide a clear picture of the distribution and spread of the diamond dimensions, which is essential for understanding their physical characteristics and variability. The IQR, in particular, helps in identifying the range within which the central 50% of the data lies, highlighting potential outliers and the overall dispersion of the data.

6.4.2 Outlier detections for 'x', 'y' and 'z' columns

- Lower thresholds for 'x' column: 1.9649999999999999

- Upper thresholds for 'x' column: 9.285

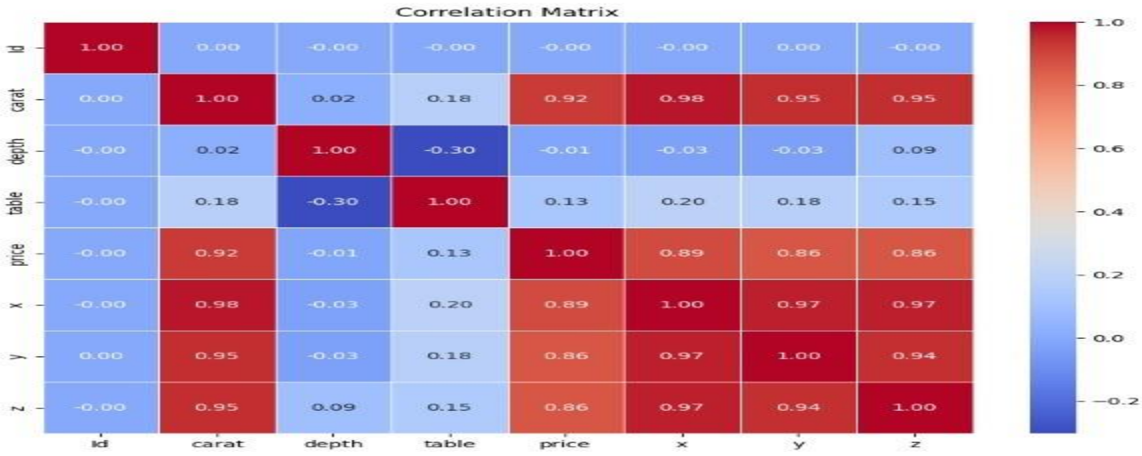
- Lower thresholds for 'y' column: 1.9899999999999993

- Upper thresholds for 'y' column: 9.27

- Lower thresholds for 'z' column: 1.2150000000000003

- Upper thresholds for 'z' column: 5.734999999999999

6.5 Visualize correlation matrix



Observations from the Correlation Matrix:

- Carat has a strong positive correlation with price, indicating that heavier diamonds tend to be more expensive.
- X, Y, and Z dimensions also show a positive correlation with price, as larger diamonds in physical dimensions are typically more valuable.
- Depth and Table have weaker correlations with price, suggesting that while they do affect diamond value, their impact is less significant compared to carat weight and physical dimensions.

7- Prepare the Data for Machine Learning Algorithms

7.1 Feature Engineering

7.1.1 creating new features: In data analytics, Feature Engineering step involves deriving new attributes from existing ones to improve predictive models. In this step, various geometric properties and ratios are calculated based on the dimensions of diamonds.

This step involves the calculation of geometric features for diamonds, including volume, diameter, density, surface area, depth percentage, and length ratios. These features offer additional insights into the physical characteristics of the diamonds, which can be valuable for analysis and modeling tasks in the domain of diamond price prediction.

7.2 Preprocess the Data

7.2.1 Encoding Categorical Features Ordinally

By ordinally encoding the categorical features 'cut', 'color', and 'clarity', this code snippet transforms them into numerical representations while preserving their ordinal relationships. This facilitates the inclusion of these features in machine learning models and analysis pipelines, contributing to more effective data-driven insights and predictions.

7.2.2 Removing Redundant Categorical Features

By removing the categorical features 'cut', 'color', and 'clarity', this code snippet streamlines the dataset for further analysis or modeling. This step is essential for focusing on the most relevant and informative features, ultimately enhancing the efficiency and effectiveness of downstream tasks.

7.3 Data Cleaning

7.3.1 Handling Outliers and Zero Values by removing them.

7.3.2 Handling duplicates values after adding new features by removing duplicate values.

8- Model Selection and Training

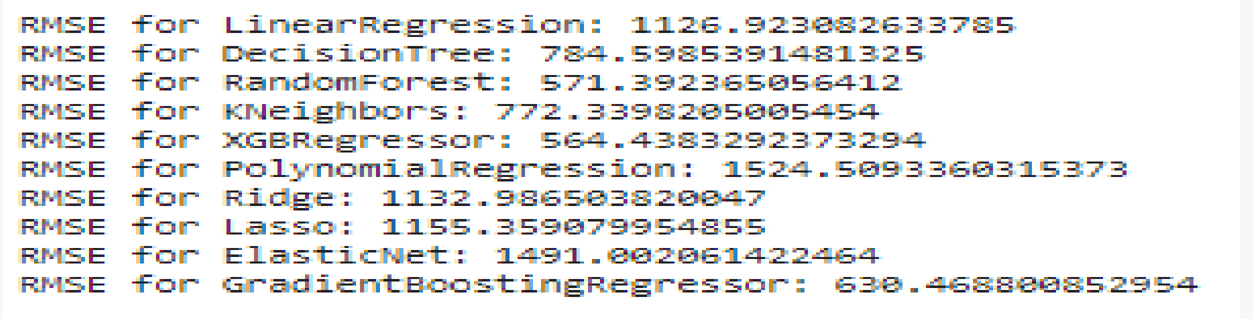
8.1 Split the Training Data

8.2 Standardizing Features using Robust Scaler

8.3 Train the Models & Evaluate the Models:

Based on the comparative analysis, it's evident that using a pipeline for the XGBRegressor model resulted in a slightly higher RMSE of 564.44. On the other hand, without employing a pipeline, the RMSE for the XGBRegressor model was slightly lower at 559.92. This indicates that the additional complexity introduced by the pipeline did not lead to significant improvements in model performance. Considering the marginally better performance achieved without a pipeline, it seems preferable to proceed without using a pipeline in this scenario. Without a pipeline, there is more flexibility to fine-tune preprocessing steps and potentially achieve better results with simpler workflows.

As in below, the best-performing model was XGBRegressor Model.



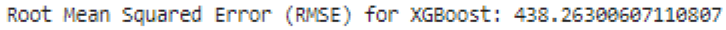
```
RMSE for LinearRegression: 1126.923082633785
RMSE for DecisionTree: 784.5985391481325
RMSE for RandomForest: 571.392365056412
RMSE for KNeighbors: 772.3398205005454
RMSE for XGBRegressor: 564.4383292373294
RMSE for PolynomialRegression: 1524.5093360315373
RMSE for Ridge: 1132.986503820047
RMSE for Lasso: 1155.359079954855
RMSE for ElasticNet: 1491.002061422464
RMSE for GradientBoostingRegressor: 630.468800852954
```

9- Fine-Tune Your Model

We applied these steps to minimize the root mean squared error (RMSE) on the training data during cross-validation.

9.1 Apply Feature Importance Analysis with XGBRegressor

9.2 Hyperparameter Tuning for XGBRegressor Using Grid Search



```
Root Mean Squared Error (RMSE) for XGBoost: 438.26300607110807
```

10- Evaluate Your System on the Test Set

10.1 Load the test dataset.

10.2 We applied feature engineering and preprocessing steps used on the training data to the test data.

10.3 Training XGBRegressor with Best Parameters and Making Predictions on the test data.

10.4 Submission Format

Prepare the submission file with the predicted prices in the format: (Id, Price).

11. Conclusion

Our journey through diamond price prediction has been thorough and insightful. Beginning with data exploration, we gained valuable insights into diamond attributes. Through meticulous preprocessing and advanced modeling techniques, particularly with the XGBRegressor, we developed accurate predictive models.

Utilizing RMSE as our guide, we fine-tuned our models for optimal performance. Applying our preprocessing steps to the test set, we confidently made predictions and prepared submissions for the competition.

In conclusion, our machine learning project not only predicts diamond prices accurately but also offers valuable insights for understanding the complexities of the diamond market.