

Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037- Concept and Technology of AI.	3- Coursework	Report Writing

Regression Analysis Report

Student Id : 2408291
Student Name : Rohit Ghimire
Section : L5CG2
Module Leader : Siman Giri
Tutor : Sunita Parajuli
Submitted on : 11-02-2025

Contents

1. Introduction:	4
Dataset Information:.....	4
Content of the Dataset(Attributes)	4
Problem Statement	4
1.3 Objective	5
2. Methodology:	5
2.1 Data Preprocessing	5
Handling Missing Values:	5
2.2 Exploratory Data Analysis (EDA).....	5
Visualizations	5
Key Insights:.....	8
2.3 Model Building	8
2.4 Model Evaluation	8
Hyper-parameter Optimization -GridSearchCV :	9
3. Conclusion:.....	9
3.1 Key Findings	9
3.2 Final Model	9
3.3 Challenges	9
3.4 Future Work	10
4. Discussion:	10
4.1 Model Performance.....	10
4.2 Impact of Hyperparameter Tuning and Feature Selection	10
4.3 Interpretation of Results	11
4.4 Limitations	11
4.5 Suggestions for Future Research	11

1. Introduction:

Dataset Information:

The dataset used for this analysis is diamond.csv, which is sourced from Kaggle - Diamond Dataset. The Dataset was created by Shivam Agrawal 8 years ago. The dataset contains 54,000+ diamonds, with the goal to predict the price of diamonds based on various features.

As we know, for any nations economic wealth is the most important thing which can be aligned with the sustainable development goals number -8 . I choose this dataset to see the rising price of diamond in dollar(US) according to its cut , carat and other features.

Content of the Dataset(Attributes)

- Price:
- Carat:
- Cut:
- Color:
- Clarity:
- x
- y:
- z:
- Table:

Statement

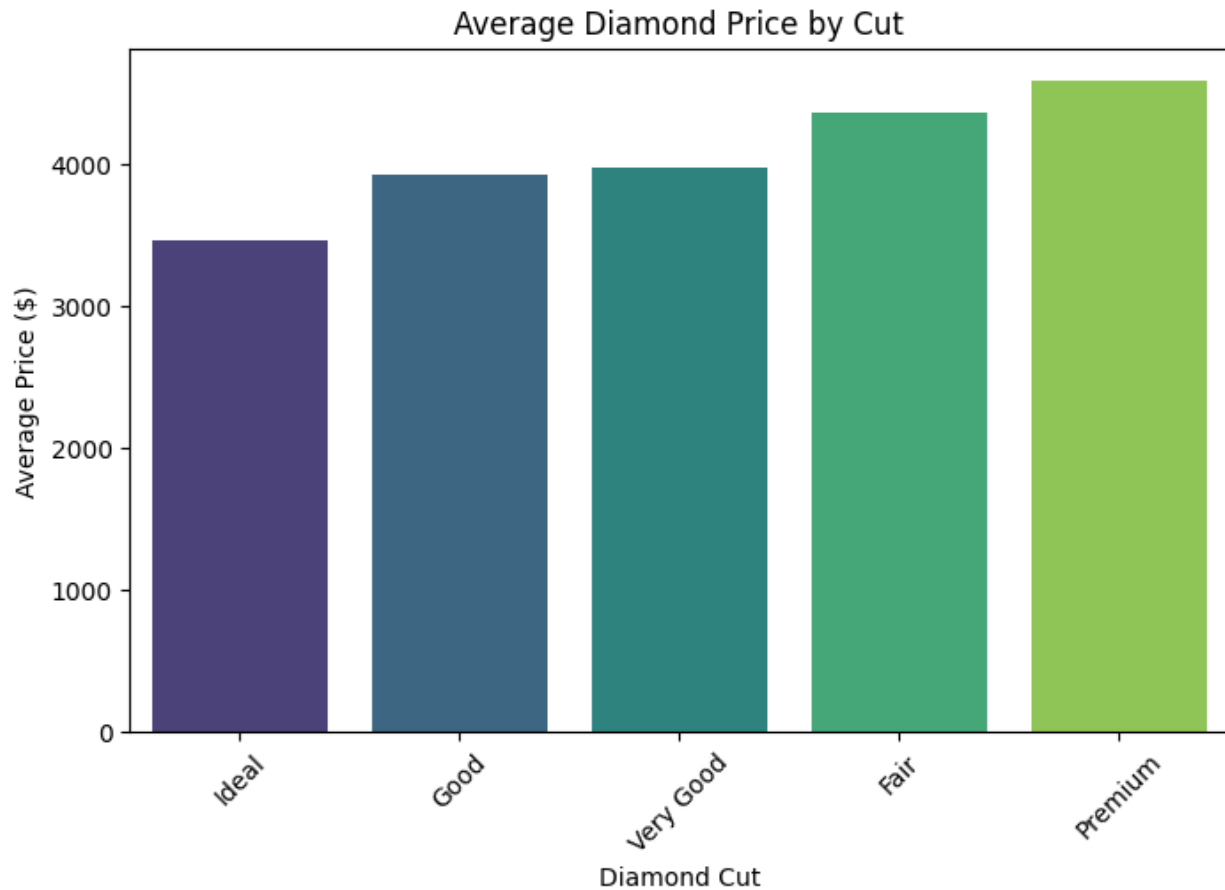
The regression model will help predict a price of diamond in dollar depending on the above attributes of dataset.

2. Methodology:

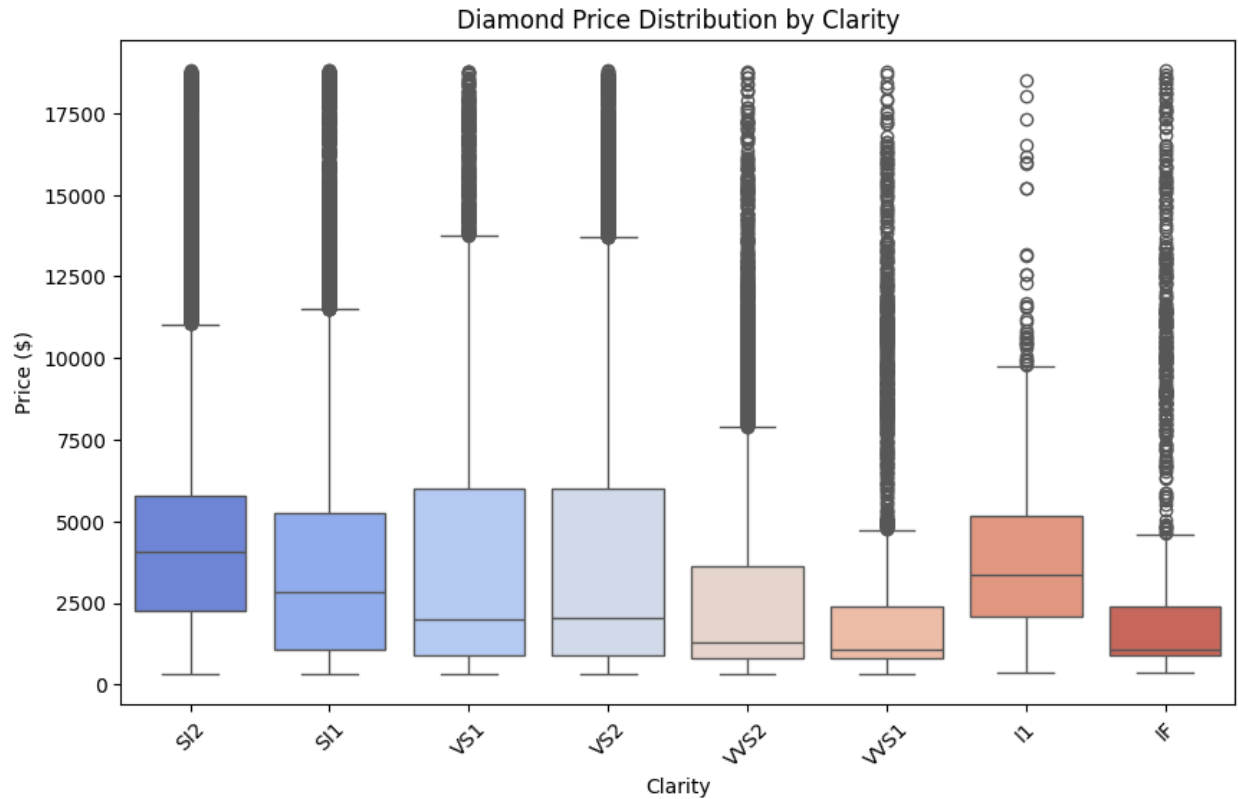
The dataset was checked for missing or inconsistent data and was cleaned as needed.

2.2 Exploratory Data Analysis (EDA)

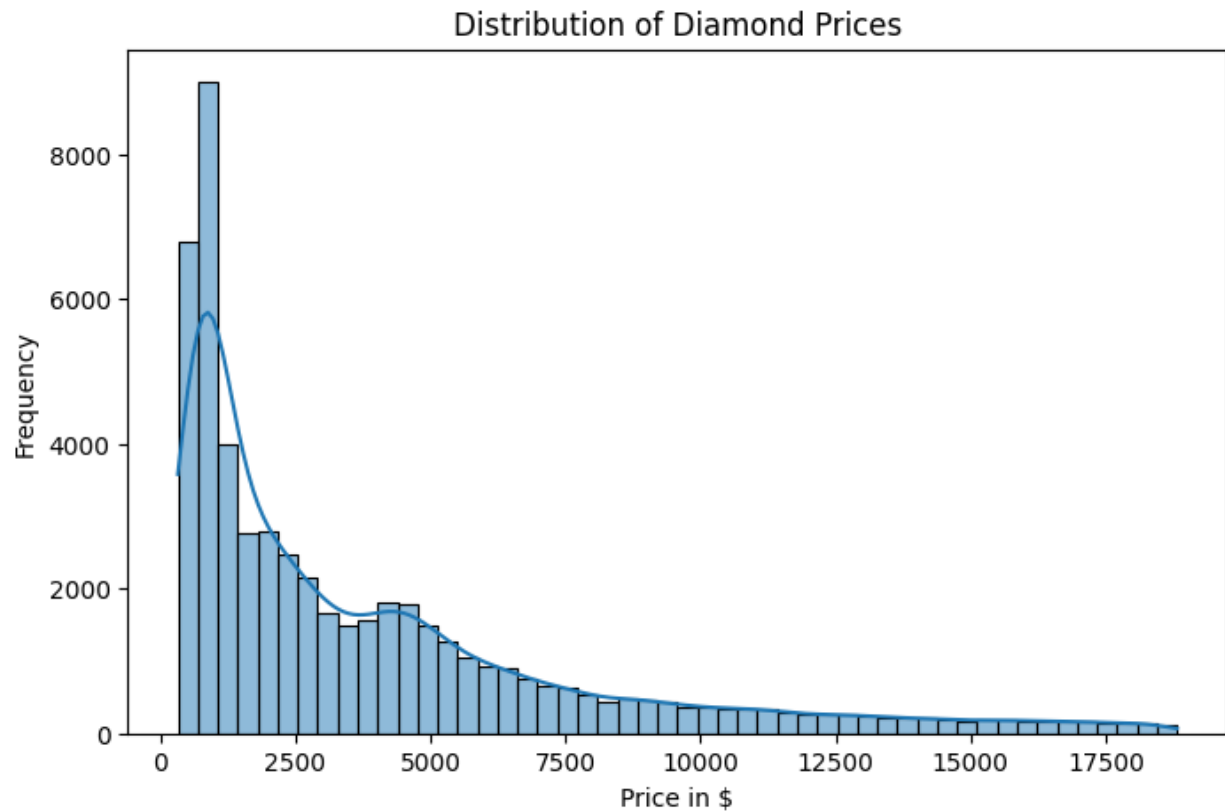
Data visualizations were done to check the average price of diamond according to cut .



The price is directly dependent on diamond cutting. The greater the critical cut of diamonds the higher the price.



The box plot explains that the Diamond price is positively correlated with clarity: higher clarity generally means higher prices. Outliers suggest that some diamonds are priced significantly higher than what might be expected based solely on their clarity. The line inside the box is median, lower is Q1(25%) of data, upper is Q3 (25%) of data and upper is(50%) of data.



Histogram in respect of price and frequency of the price of diamonds.

Key Insights:

For example, the carat variable has a strong correlation with the price of diamonds.

2.3 Model Building

Model was build by splitting the dataset into train test split and use the linear regression from scratch and were calculated to predict the price.

2.4 Model Evaluation

The model's performance was evaluated using the following metrics:

- Root Mean Squared Error (RMSE):
- R-squared:
- Mean Squared Error (MSE)

Hyper-parameter Optimization -GridSearchCV:

It was used to optimize hyperparameters and improve model performance. The optimal parameters were found based on cross-validation.

▲ 3. Conclusions:

3.1 Findings

The model's performance on the test dataset was evaluated using R-squared and MSE. The results showed that the Linear Regression model performed well, with a good fit and reasonable prediction errors.

RMSE: 5469.985955759538

MSE: 29920746.35620659

MAE: 4013.3260353062324

R² Score: 0.9206226715690513

3.2 Final Model

The Linear Regression model was chosen as the most effective for predicting diamond prices. The final model achieved an R-squared value of 0.92 and an MSE of 1558.

3.3 Challenges

Some challenges encountered during the project included handling categorical variables and the large dataset size, which required significant computational power.

Future improvements could be using more complex models like Gradient Boosting , which may give better performance. Additional feature engineering could also improve accuracy.

4. Discussion:

4.1 Model Performance

The model's performance was evaluated using R-squared and MSE, and it was found that the model performed well on the test data.

Linear Regression:

Mean Squared Error (MSE): 1238015.3387700708

Mean Absolute Error (MAE): 734.0152467397219

R-squared (R²): 0.9206195019546656

Random Forest Regression:

Mean Squared Error (MSE): 1558.0443567235197

Mean Absolute Error (MAE): 3.468673217154869

R-squared (R²): 0.9999000995115809

4.3 Interpretation of Results

The key features contributing most to the model's predictions were carat and cut. This indicates that larger carats and better cut quality are strongly correlated with higher diamond prices.

▲ 4.4 Limitations

Despite the successful modeling, some limitations remain, including potential overfitting and the inability to capture very high-end diamonds accurately due to the limitations of the dataset.

4.5 Suggestions for Future Research

Experimenting with advanced regression algorithms like Lasso or Ridge Regression to reduce overfitting.

Decreasing the dataset size to include more diamonds and different types of data to improve the model's robustness.

Implementing feature engineering to create new features that better capture the relationship between diamond features and price.