UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

| 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

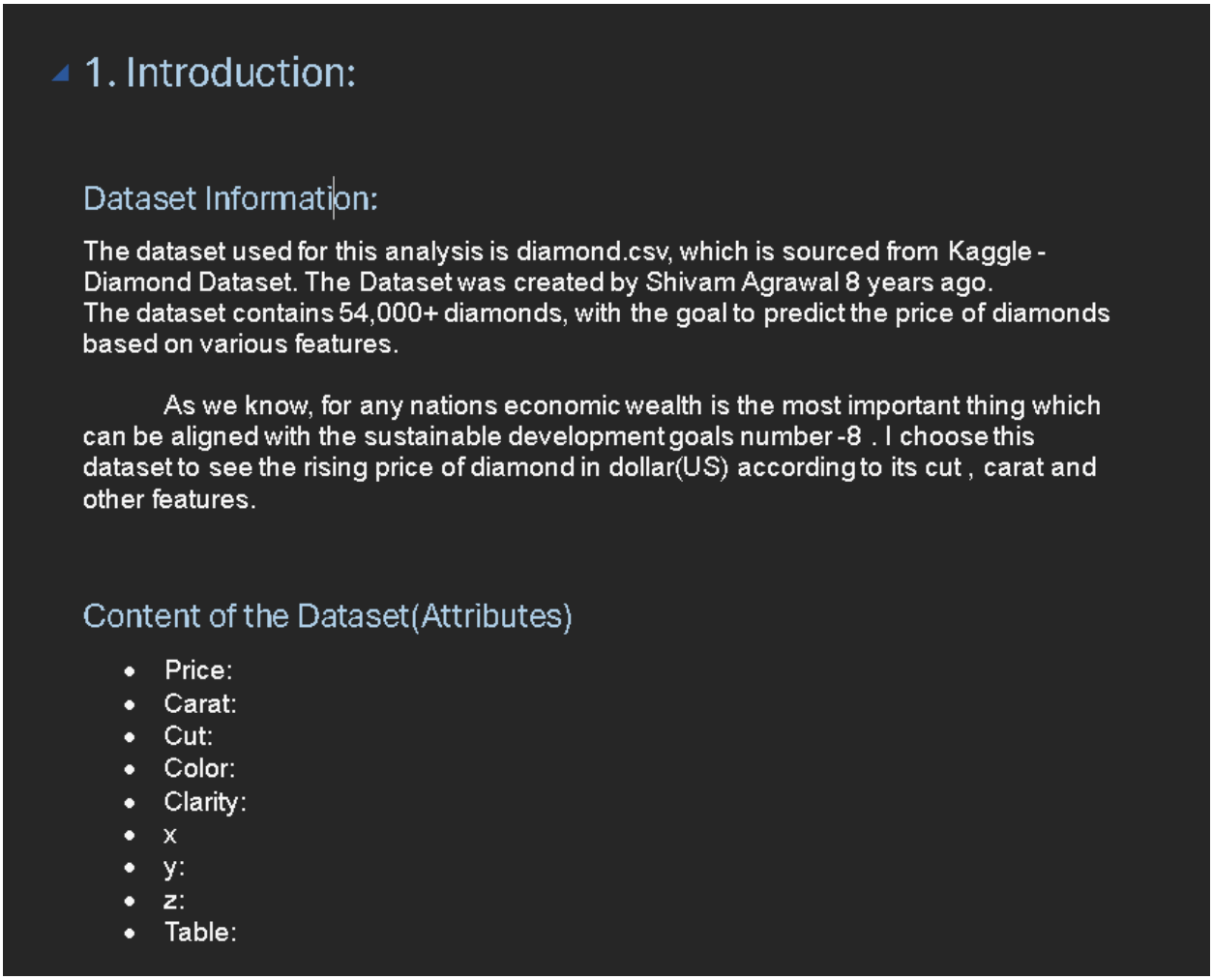
Section : L5CG2

Module Leader : Siman Giri

Tutor : Sunita Parajuli


Submitted on : 11-02-2025

## Contents




### Statement

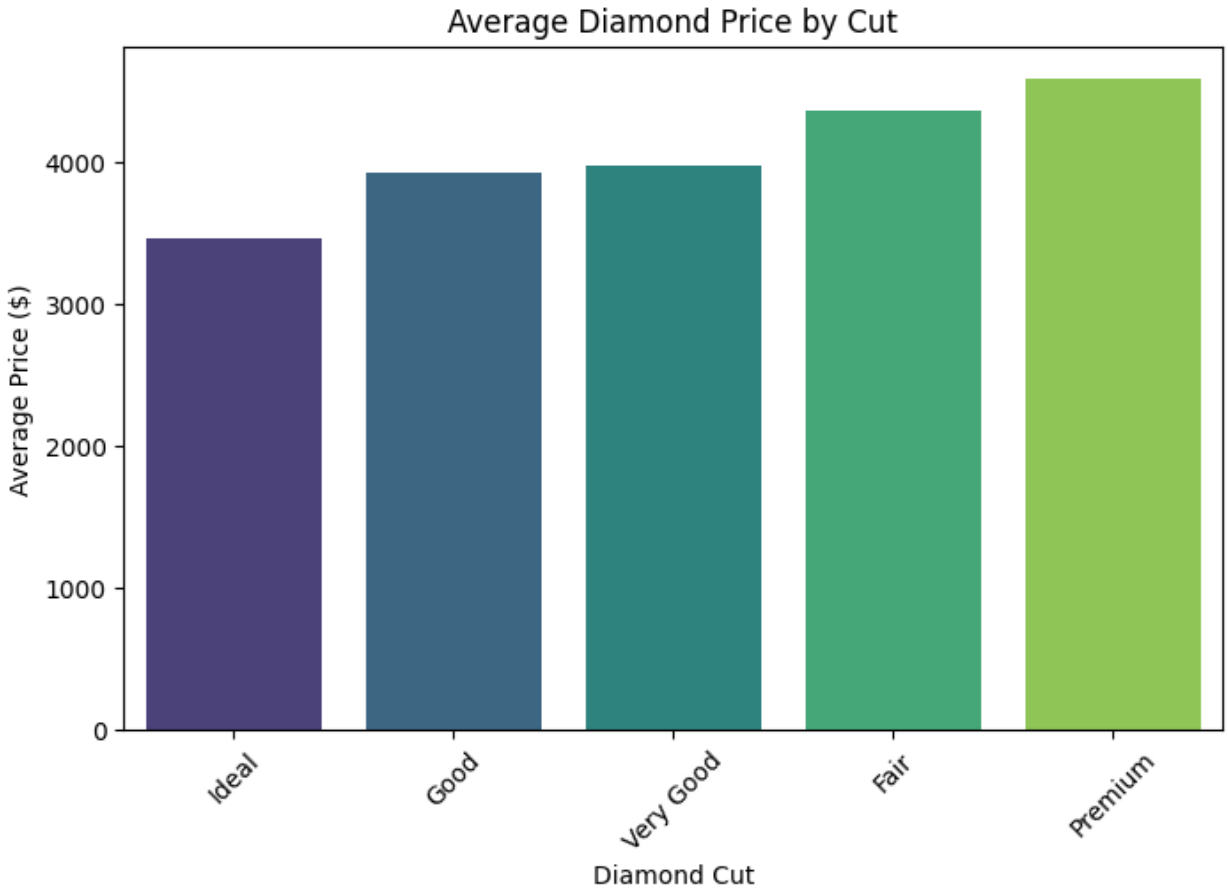
The regression model will help predict a price of diamond in dollar depending on the above attributes of dataset.

## 2. Methodology:

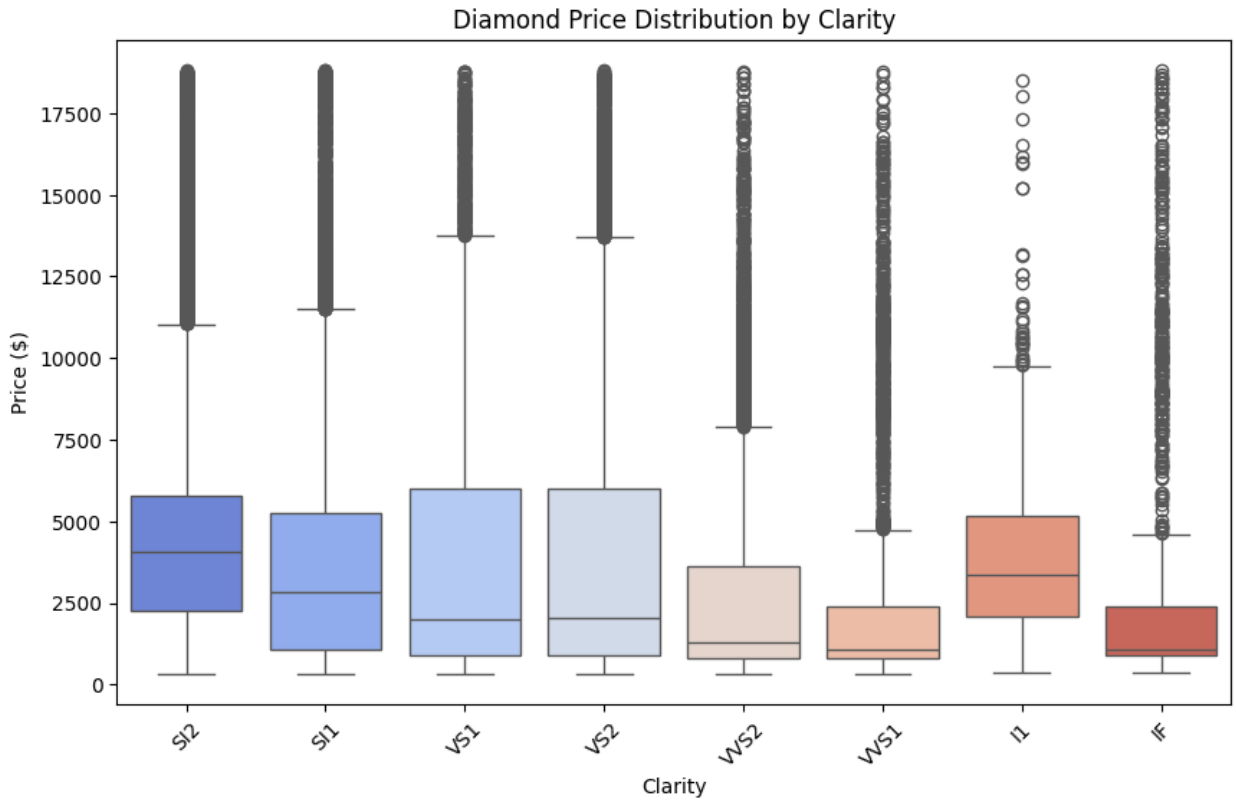
The dataset was  checked for missing or inconsistent data and was cleaned as needed.

## 2.2 Exploratory Data Analysis (EDA)

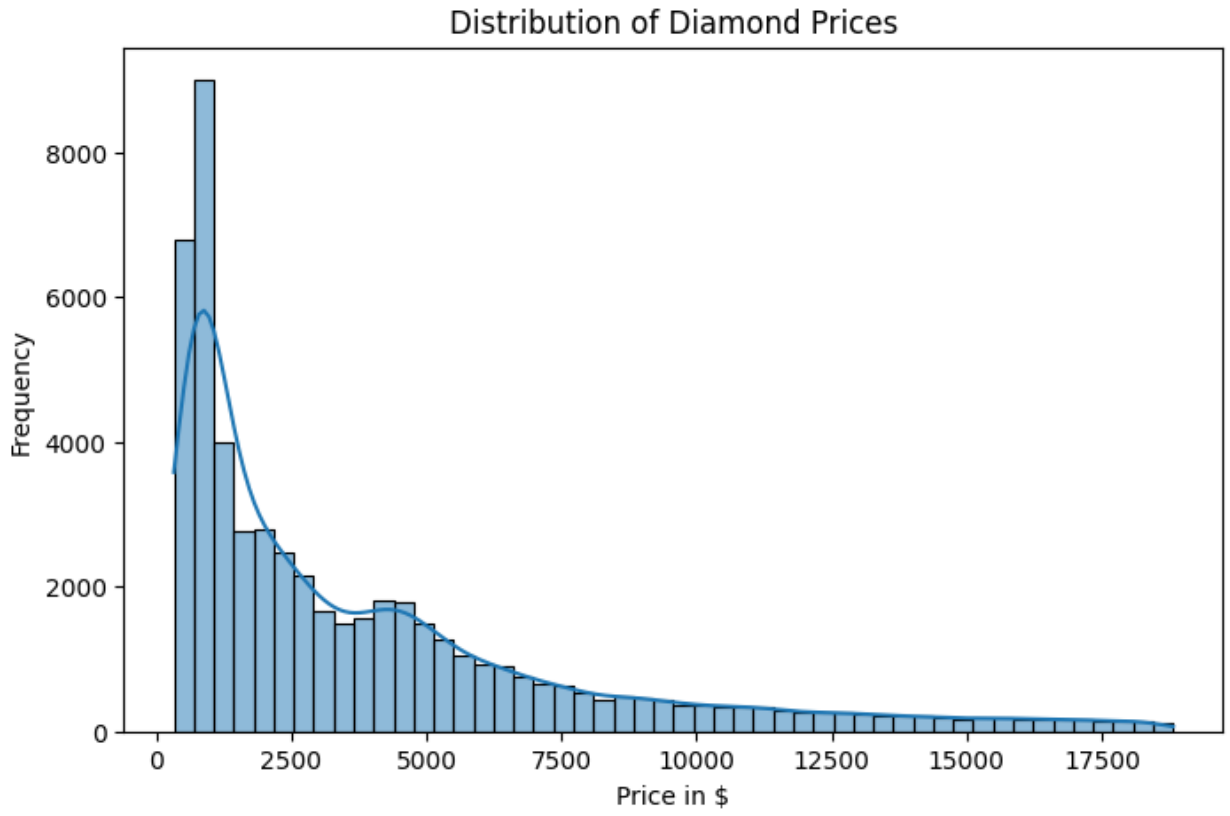
Data visualizations were done to check the average price of diamond according to cut .




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.



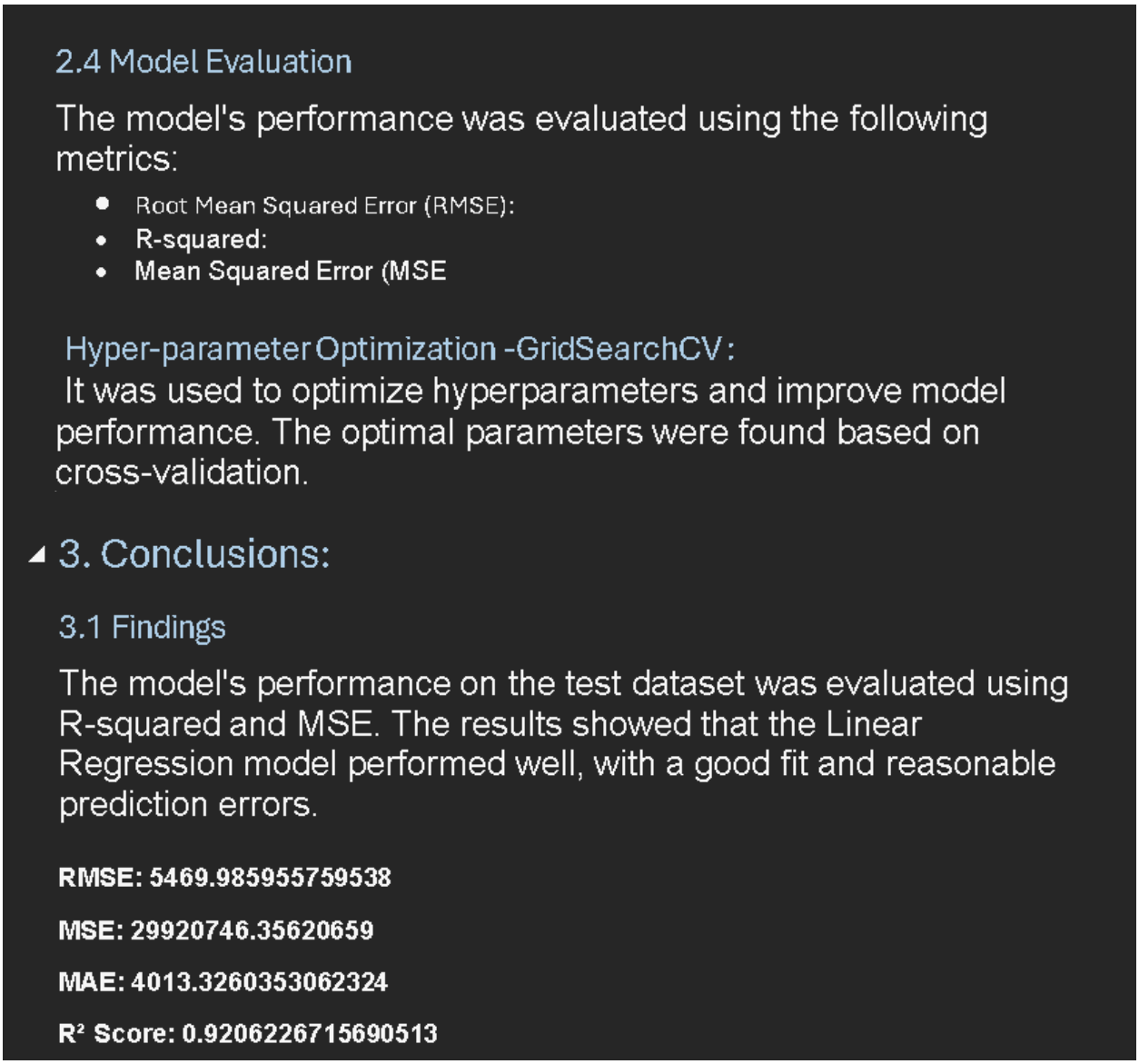

Histogram in respect of price and frequency of the price of diamonds.

### 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.



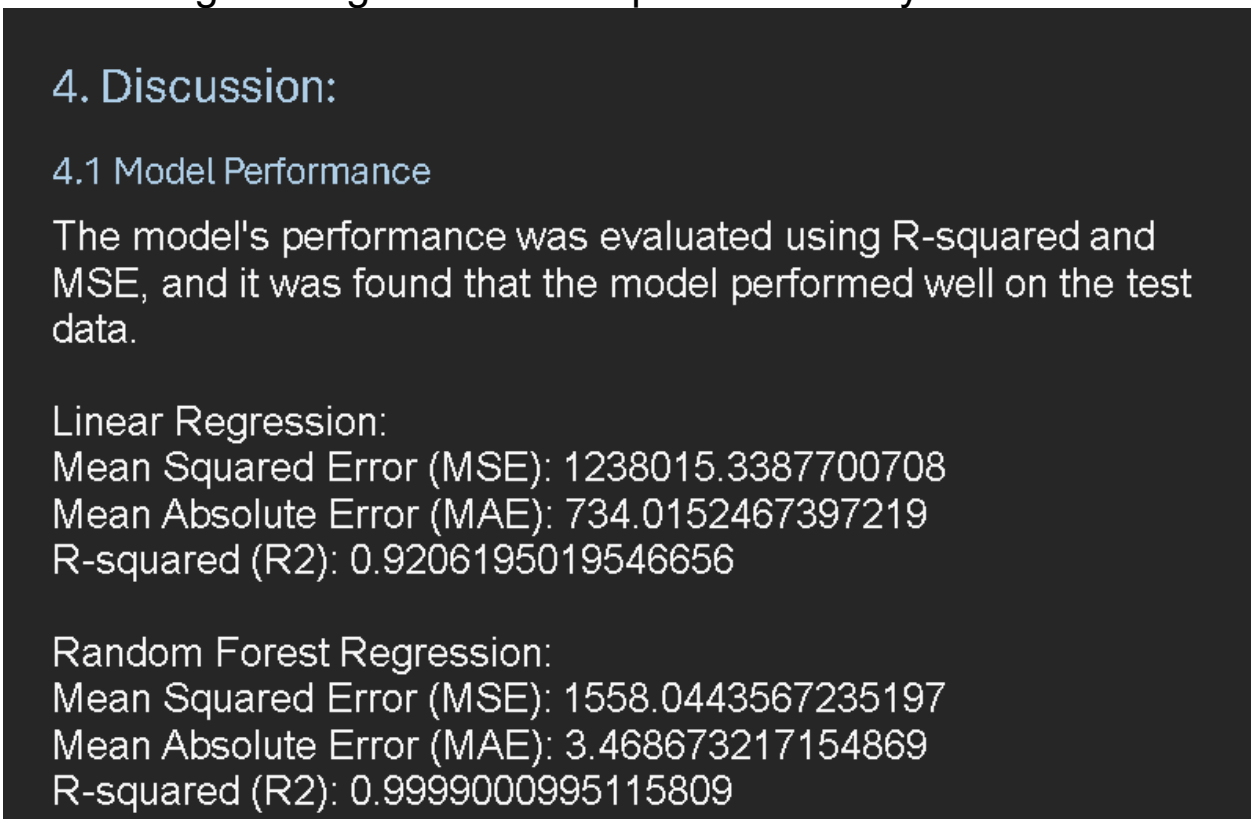

## 3.2 Final Model

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.




### 4.3 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.