Laboratory assignment

Component Diamond Dataset

Authors: Murariu Tudor Cristian and Vâtcă Cristina
Group: 12

December 6, 2024

Contents

1 Introduction

This report analyzes the Diamond Prices dataset to solve two machine learning tasks:

1. Predicting the cut of diamonds (classification).
2. Predicting the price of diamonds (regression).

The dataset contains various features such as carat, cut, color, clarity, and dimensions (x, y, z), providing a comprehensive basis for analysis.

2 Task 1: Predicting Cut (Classification)

2.1 Problem Definition

The goal is to classify the quality of the diamond's cut into one of five categories: *Ideal*, *Premium*, *Very Good*, *Good*, and *Fair*.

2.2 Problem Specification

Input Data:

- Features such as carat, color, clarity, depth, and table.

Output:

- Predicted cut category.

Preconditions:

- Dataset includes labeled examples for all cut categories.

Postconditions:

- The model accurately predicts the cut for unseen data.

2.3 Learning Task Specification

- Task (T): Classify each instance into one of the five cut categories.
- Performance (P): Evaluation metrics include accuracy, precision, recall, and F1-score.
- Experience (E): The model is trained using supervised learning methods on labeled data.

3 Task 2: Predicting Price (Regression)

3.1 Problem Definition

The aim is to predict the price of a diamond based on its features.

3.2 Problem Specification

Input Data:

- Features such as carat, cut, color, clarity, depth, and table.

Output:

- Predicted price (continuous variable).

Preconditions:

- Dataset includes sufficient data covering all price ranges.

Postconditions:

- The model generalizes well to unseen data, providing accurate predictions.

3.3 Learning Task Specification

- Task (T): Predict the price of diamonds.
- Performance (P): Metrics include Mean Squared Error (MSE) and Mean Absolute Error (MAE).
- Experience (E): The model is trained on historical data using supervised learning methods.

4 Dataset Description

Each record is a random diamond with its own characteristics.
The dataset consists of 10 attributes:

- **Carat:** Weight of the diamond.
- **Cut:** Quality of the cut.
- **Color:** Diamond color, graded from D (best) to J (worst).
- **Clarity:** Measure of diamond imperfections.
- **Depth:** Total depth percentage.
- **Table:** Width of the top of the diamond relative to its widest point.
- **Price:** Price in US dollars.
- **x, y, z:** Dimensions of the diamond in mm.

5 Feature Analysis

	carat	depth	table	price	x	y	z
count	53940.00	53940.00	53940.00	53940.00	53940.00	53940.00	53940.00
mean	0.80	61.75	57.46	3932.80	5.73	5.73	3.54
std	0.47	1.43	2.23	3989.44	1.12	1.14	0.71
min	0.20	43.00	43.00	326.00	0.00	0.00	0.00
25%	0.40	61.00	56.00	950.00	4.71	4.72	2.91
50%	0.70	61.80	57.00	2401.00	5.70	5.71	3.53
75%	1.04	62.50	59.00	5324.25	6.54	6.54	4.04
max	5.01	79.00	95.00	18823.00	10.74	58.90	31.80

5.1 Correlation Heatmap

There are no missing values in any of the fields since the count for each field equals the total number of entries in the dataset. The data appears to be equally distributed with little skewness and few outliers, as indicated by the mean values being near the 50th percentile (median). Each field's minimum and maximum observations are represented by the Min and Max values, which fit within acceptable ranges (for example, for year, month, etc.). Furthermore, the majority of the data points are near the mean and show little volatility, as indicated by the comparatively small standard deviation for all fields.

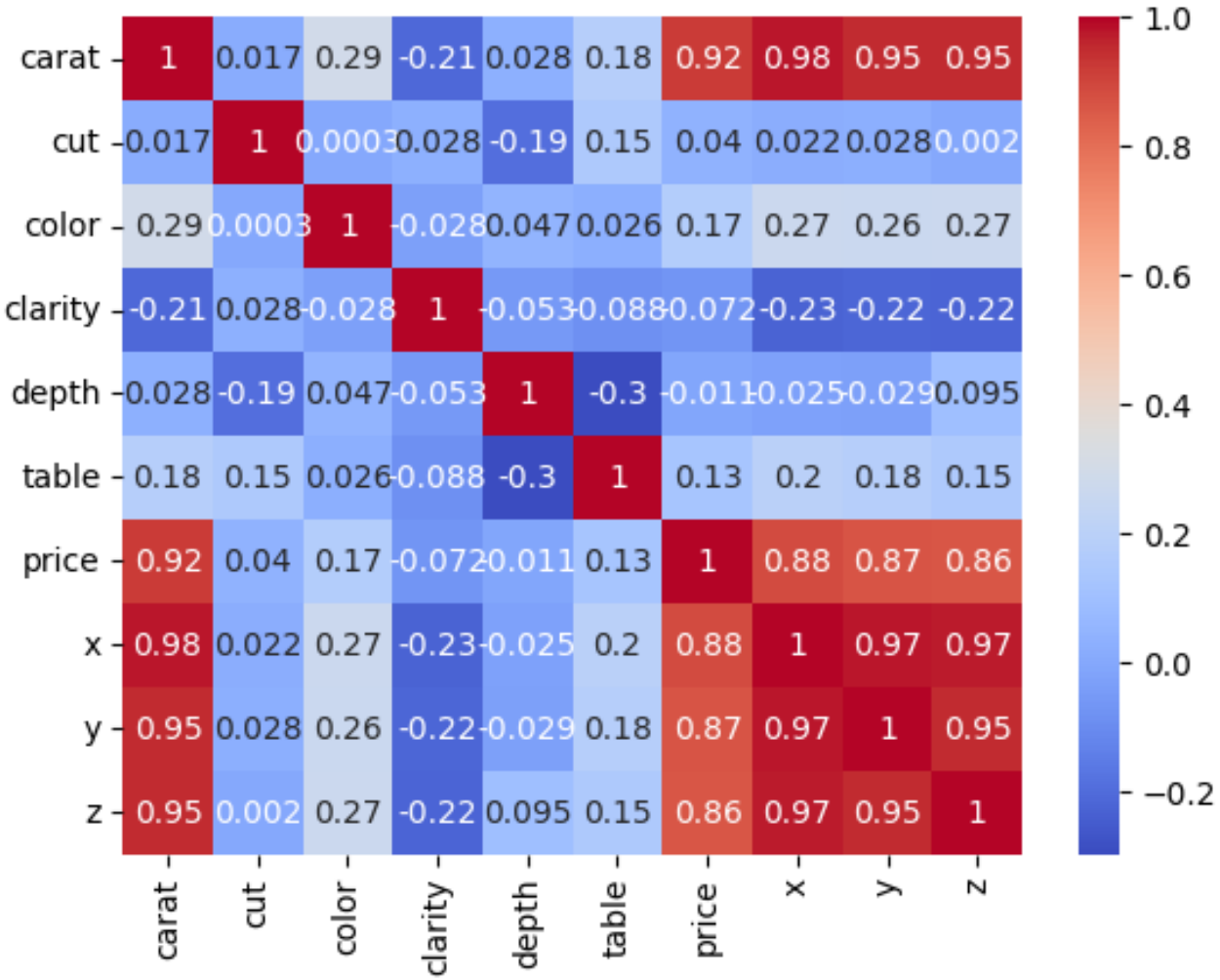

Figure 1: Heatmap of Feature Correlations.

We can see how each field correlates with every other field in the heatmap above. For example it seems that the fields: carat, price, x, y and z are strongly connected.

5.2 Data Distribution

The distribution of each feature in the dataset is depicted in the plot below.

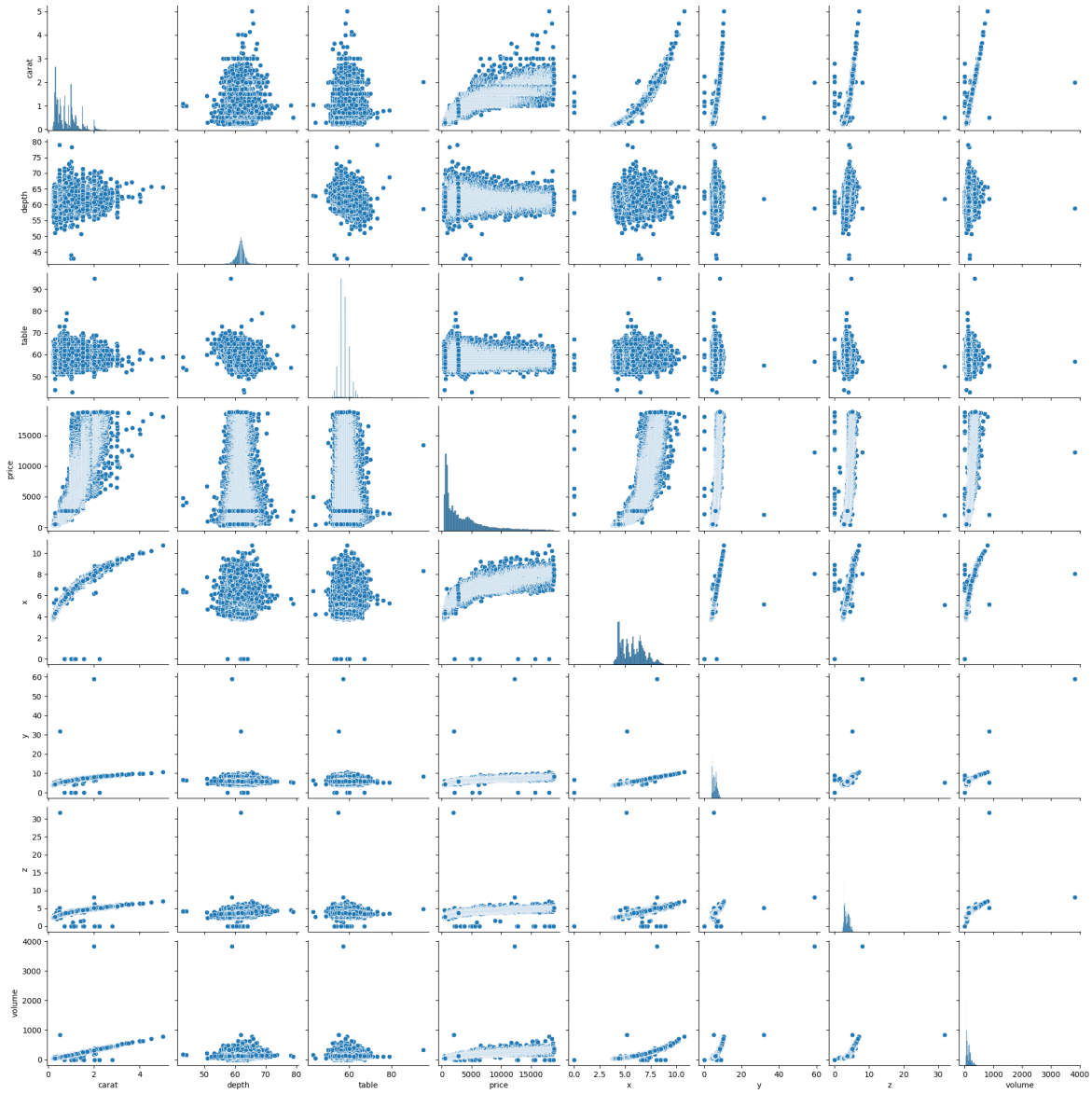

Figure 2: Data Distribution

I would like to also show the count distributions for the data we will to predict:

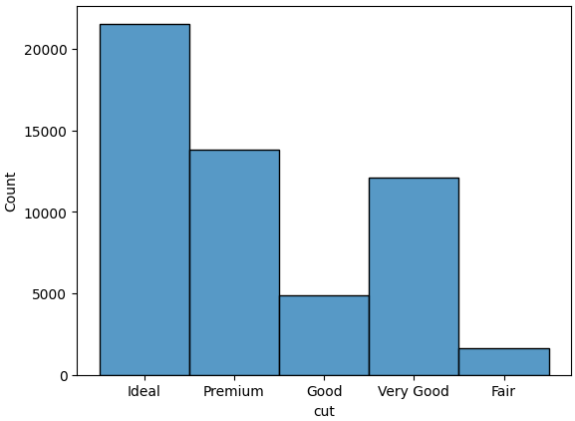

Figure 3: Number of daimons per Cut

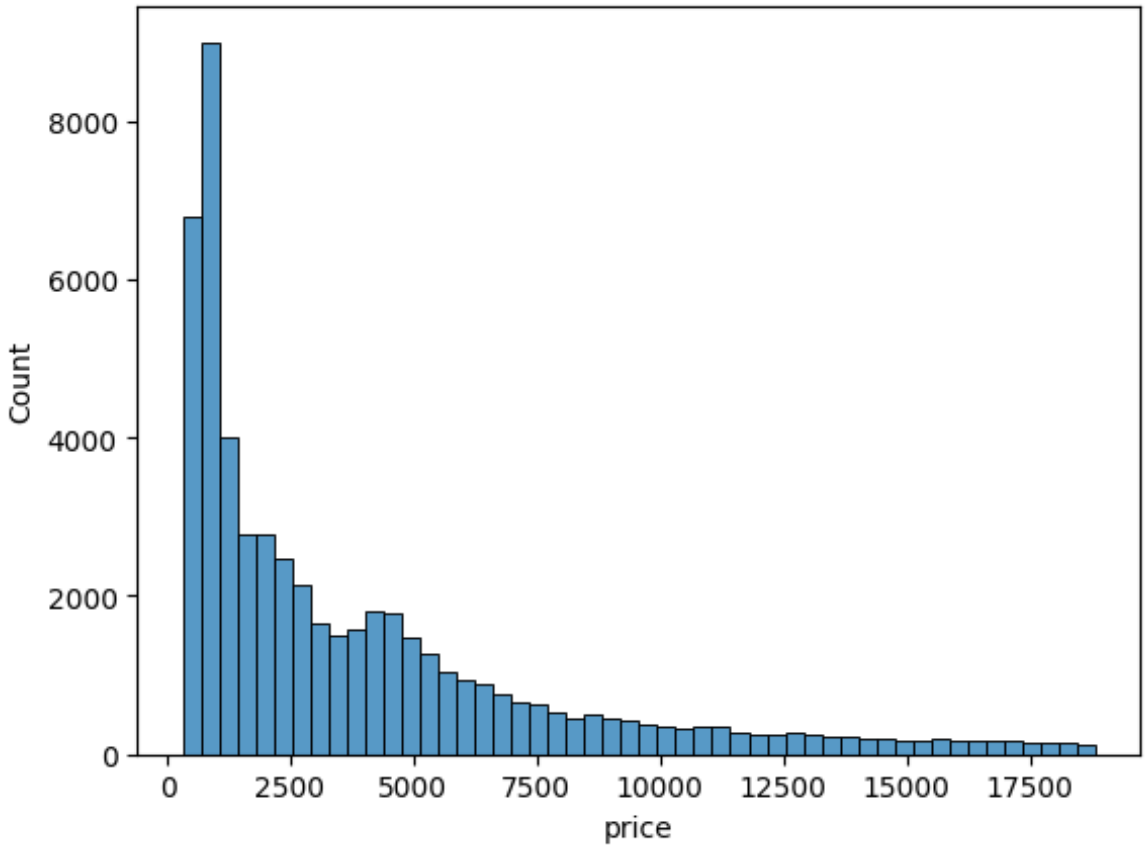

Figure 4: Number of daimons for each range of prices

5.3 Feature Importance

- For predicting cut: price, carat, clarity and table are the most significant.

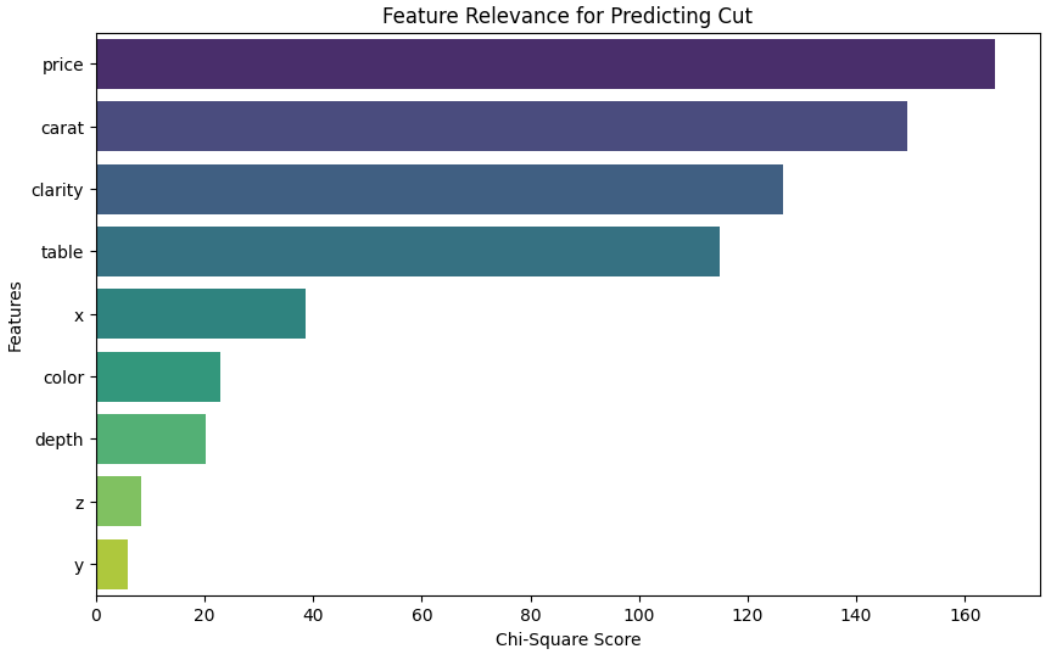

Figure 5: Corelations with the Cut

- For predicting price: carat, x, y and z hold the highest importance.

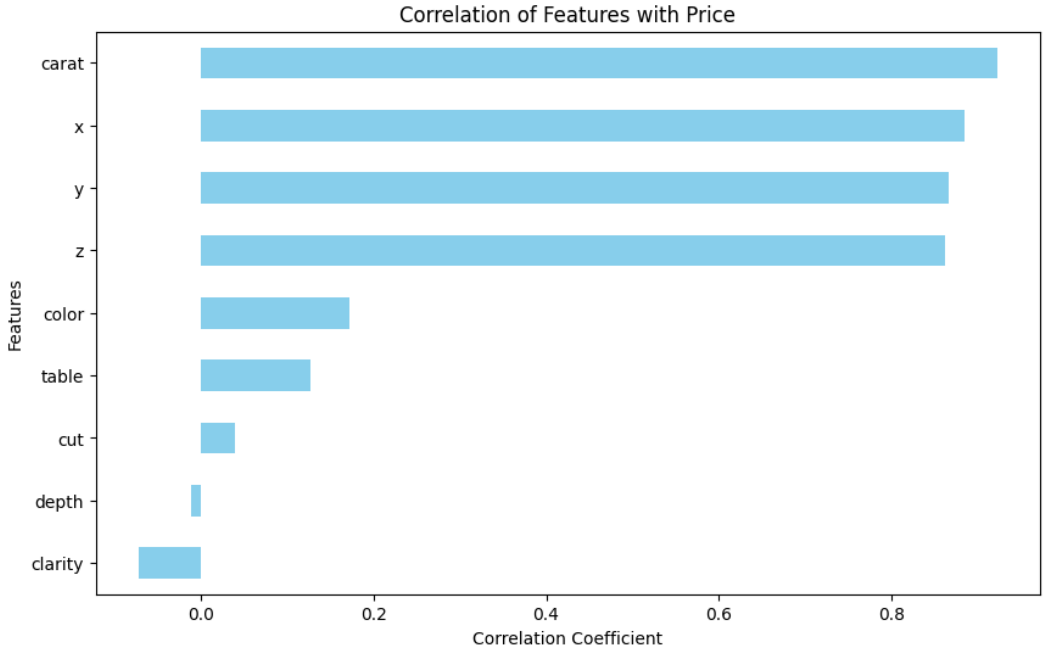

Figure 6: Corelations with the Price