STAT40830 - Homework 1

Shreemadhi Babu Rajendra Prasad

Introduction

The **'diamonds'** is an in built dataset from **'ggplot2'** package which contains *pricing and quality* information of around 50,000 diamonds.

Each row represents the data of a single diamond with the key variables:

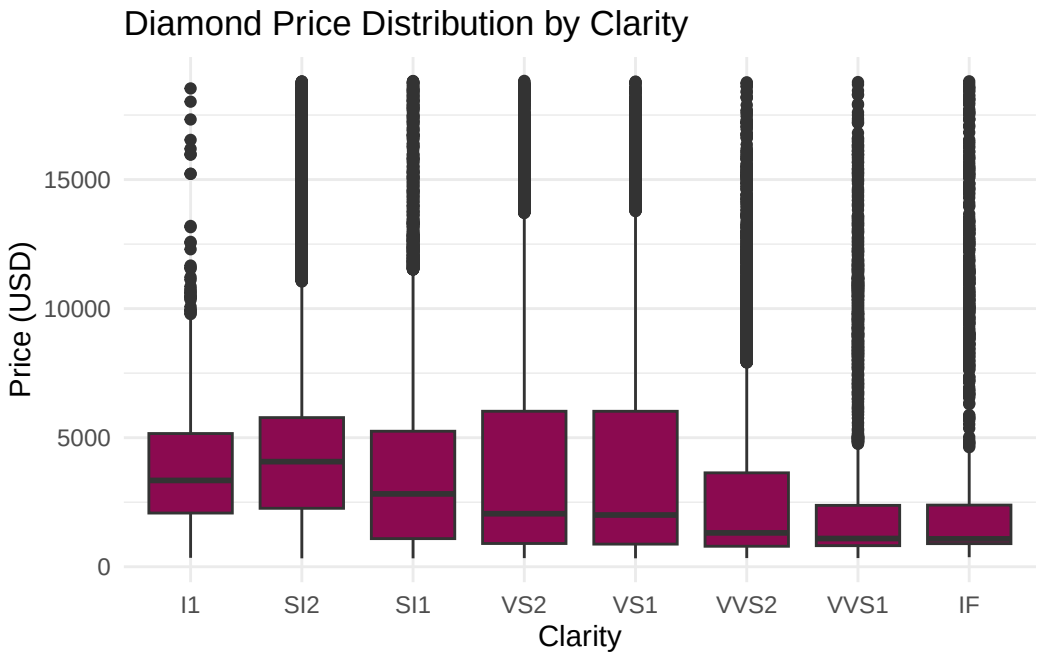
- **carat**: *Weight of the diamond*
- **cut**: *Quality of the cut (Fair, Good, Very Good, Premium, Ideal)*
- **color**: *Diamond color, from D (best) to J (worst)*
- **clarity**: *Measurement of internal flaws*
- **depth**: *Total depth percentage = z / mean(x, y)*
- **table**: *Width of the top of the diamond relative to its widest point*
- **price**: *Price in US dollars*
- **x, y, z**: *Length, width, and depth (in mm)*

This dataset is commonly used for demonstrate **data visualization, statistical modeling, etc.**

```
[1] 53940    10
```

The dataset consists of **53940 rows** and **10 columns**.

Diamond Price Distribution by Clarity



This **boxplot** shows a variation between *diamond prices* across different **clarity levels**.

- The median price differs across the **clarity types**.
- Clarity grades like IF and VVS1 shows a **lower median prices**, despite being higher quality.
- Diamonds with lower clarity (e.g., I1, SI2) often show **higher median prices** due to larger carat sizes.
- There are **many outliers** — especially in lower clarity grades which indicates that **price is influenced by multiple factors**, not just clarity.