STAT40830 - Homework 1

Alan McLoughlin - 23212461

May 31, 2025

Dataset Chosen - diamonds

Overview

For the purpose of this analysis, I will use the **Diamonds** dataset, which is a dataset that comes built-in with the ggplot2 package. This dataset contains information about *53,940* round-cut diamonds, with *10* variables measuring various pieces of information about the diamonds.

Understanding Dataset

Table 1: First 5 Observations of the Diamonds Dataset

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75

As shown in Table 1, the diamonds dataset has the following 10 variables:

- **carat -** Weight of diamond.
- **cut -** Quality of the cut.
- **color -** Colour of the diamond.
- **clarity -** Measurement of how clear diamond is.
- **depth -** Total depth percentage.
- **table -** Width of top of diamond relative to widest point.
- **price -** Price ($).

- **x -** Length (mm).
- **y -** Width (mm).
- **z -** Depth (mm).

Descriptive Statistics

In plot section, the numeric variables *price*, x, y and z are used. The below output shows summary stats for each of these variables.

```
     price             x                y                z         
 Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800  
```

The average price is $3,933, and the mean is greater than median, suggesting price is positively skewed. The middle 50% of diamonds are priced between $950 and $5,324. Looking at length (x), width (y) and depth (z) of diamonds, all appear to be fairly normally distributed as mean is almost equal median. Depth appears to be the smallest dimension on average, as the middle 50% of values lie between 2.910mm and 4.040mm, which are both lower than for length and width. However, the max of depth is greater than length, suggesting depth has some significant outliers. The max value of width is also large, suggesting this measurement also has extreme outliers.

Plots

Average price per cut

Figure 1 shows the average price of a diamond for each cut (*Fair, Good, Very Good, Premium, Ideal*). To create this, a table was created storing the average price per cut, and then the bar chart was plotted using ggplot2.

Looking at the output, the Premium cut has the highest average price ($4,584), followed closely by fair ($4,359). The lowest average price was seen for the ideal cut ($3,458).

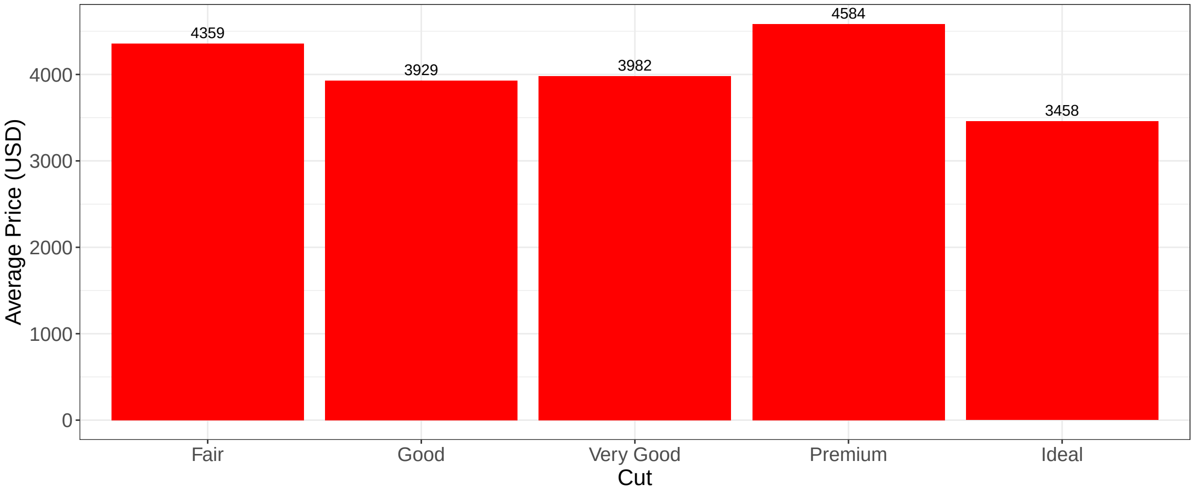

Figure 1: Average diamond price per cut

Boxplots of length (x), width (y) and depth (z)

Figure 2 shows boxplots for the variables length (x), width (y) and depth (z). Boxplots were created using ggplot2, and plots were placed size by side using gridExtra.

Looking at output produced, all three dimensions look to be fairly normally distributed. Width (y) and depth (z) appear to have some extreme outliers, for example one diamond has a width of just under 60mm, while another has a width over 30mm. All have a value of 0mm, which suggests data input error or really small diamonds. Overall, length appears to be the most stable measurement, with most values falling within bounds of (Q1−1.5×IQR, Q1+1.5×IQR).

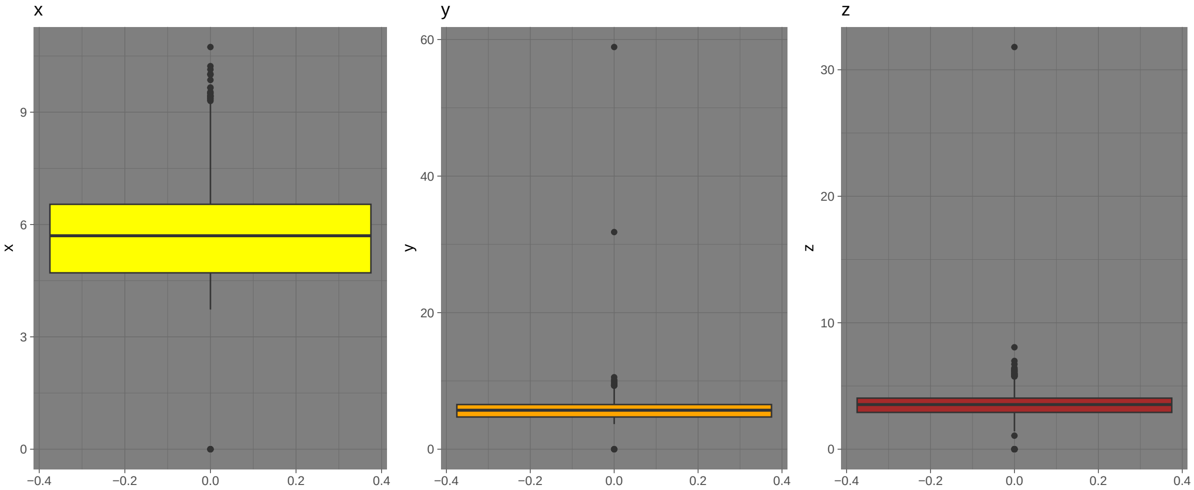

Figure 2: Boxplots of length (x), width (y) and depth (z)