Statistical Data Management

Project 1

By:

Nathan Wemmer and Naomi Burhoe

Instructor:

Jun Ye

1. **Introduction (goal, summation of analysis)**

The data analyzed in this report was obtained from a dataset that includes 308 round-cut diamonds. The data initially appeared in a newspaper advertisement and was accessed for this report via the Journal of Statistics Education data archive.

1. **Explanations of the data**

The ‘IDNO’ variable is just an identification number that is assigned for each data point. These were adjusted when data was randomly chosen from the dataset.

The ‘WEIGHT’ is the weight of each diamond, in carats.

The ‘COLOR’ attribute is represented by a letter of the alphabet, with ‘D’ being the top color purity grade. The purity grade lessens the further in the alphabet you get.

The ‘CLARITY’ variable represents the diamonds clarity grade. It is represented by one of five grades. The abbreviations along with their meanings are as follows, in descending order: “IF”, meaning internally flawless, ‘VVS1 and VVS2’, meaning very, very slightly imperfect, and ‘VS1 and VS2’, meaning very slightly imperfect.

The ‘RATER’ variable represents the location the diamond was evaluated at. The possible values are ‘GIA’, ‘IGI’, ‘HRD’ meaning Gemmological Institute of America, International Gemological Institute, and Hoge Raad Voor Diamant, respectively.

The last variable is PRICE, and it is the price of each diamond rated, in Singapore dollars. ‘WEIGHT’ and ‘PRICE’ are continuous variables while COLOR, CLARITY, and RATER are categorical variables.

1. **Details of the methods**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | WEIGHT | PRICE |  | COLOR | CLARITY | RATER |
| Min | 0.18 | 823 |  | D: 6 | IF: 5 | GIA: 99 |
| 1st Quartile | 0.5 | 3271 |  | E: 17 | VS1:40 | IGI: 1 |
| Median | 0.635 | 4580 |  | F:28 | VS2:18 |  |
| Mean | 0.6523 | 4986 |  | G: 17 | VVS1: 12 |  |
| 3rd Quartile | 0.8 | 6407 |  | H: 19 | VVS2: 25 |  |
| Max | 1.1 | 15582 |  | I:13 |  |  |

**L.O.S.S of each variable**

|  |  |
| --- | --- |
| Variable: **Weight**  (L)ocation: Mean: **0.6523**, Median: **0.635**  (O)utliers: **None**  (S)hape: **Approximately Normal**  (S)pread: Standard Deviation: **0.2291171** | Variable: **Price**  (L)ocation: Mean: **4986** Median: **4580**  (O)utliers:**15582, 11419**  (S)hape: **Skewed right**  (S)pread: Standard Deviation: **2861.012** |
| Variable: **Color**  (L)ocation: Mode: **H** - 19/100 - **19%**  (O)utliers: **NA**  (S)hape: **NA**  (S)pread **D-6% E-17% F-28% G-17% H-19% I-13%** | Variable: **Clarity**  (L)ocation: Mode: **VS1** - 40/100 - **40%**  (O)utliers:**NA**  (S)hape: **NA**  (S)pread:**IF-5% VS1-40% VS2-18% VVS1-12% VVS2-25%** |
| Variable: Rater  (L)ocation: Mode: **GIA** - 99/100 - **99%**  (O)utliers: **NA**  (S)hape: **NA**  (S)pread: **GIA-99% IGI-1%** |  |

1. **Results of the data analysis (graphs & tables)**

|  |  |
| --- | --- |
| **Figure 1** | **Figure 2** |
|  |  |
| **Figure 3** | **Figure 4** |
| **Figure 5** | **Figure 6** |
| **Figure 7** | **Figure 8** |
| **Figure 9** | **Figure 10** |

|  |  |
| --- | --- |
| **Figure** | **Explanation** |
| 1 | Shows relative frequencies for the color attribute |
| 2 | Shows relative frequencies for the clarity attribute |
| 3 | Shows how the color attribute is distributed (in a different way). |
| 4 | Shows how the diamond prices vary when categorized by color   * Some price outliers are present in the distributions for colors E, F, and G |
| 5 | Shows how the diamond weights vary when categorized by color   * Some weight outliers are present in the distributions for colors E, F, G, and I |
| 6 | Shows how the clarity attribute is distributed (in a different way) |
| 7 | Shows how the diamond prices vary when categorized by clarity   * Some price outliers are present in the distributions for clarity ratings VVS1 and VVS2 |
| 8 | Shows how the diamond weights vary when categorized by clarity   * Some weight outliers are present in the distributions for clarity ratings F and VVS1 |
| 9 | Shows how diamond prices are distributed   * As one can see in this figure, there appears to be at least one outlier present. This outlier may be contributing to the skewness of the histogram |
| 10 | Shows how diamond weights are distributed   * This figure shows an approximately normal distribution. |

1. **Discussions and conclusions/summary (thoughts about results)**

Since we trimmed our dataset down to 100 observations, we deemed one of the variables to be unnecessary for use any comparisons. The ‘RATER’ variable, which specifies where a diamond was rated, became relatively useless for comparison due to the fact that our data included 99 observations at the Gemmological Institute of America and only 1 observation at the International Gemmological Institute. As, it would be difficult to draw any important conclusions from that variable.

1. **References (link, paper, book)**

Data source: <https://www2.stetson.edu/~jrasp/data.htm>

* Scroll down and click ‘DIAMONDS.XLS’

Code References:

* The R Book
* [Pie Chart Setup](https://chemicalstatistician.wordpress.com/2015/02/03/how-to-get-the-frequency-table-of-a-categorical-variable-as-a-data-frame-in-r/)
* [Colors for Plots](http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf)