# Carat Writeup

Humaid Billoo

2025-03-14

# 1 Data Description

The data set, `diamonds4.csv` , contains 5 variables that describes more than 1000 different diamonds for sale.

- `Carat` : Measures a diamond's weight, not its size, the heavier the diamond the rarer it is.
- `Clarity` : *fill in description*
- `Color` : How colorless a diamond is, the more colorless the diamonds, the rarer it tends to be
- `Cut` : *fill in description*
- `Price` : the price value of the diamond in USD

## 1.1 Variable Analysis: Carat

The diamond Carat refers to the weight of a diamond. From the Blue Nile (https://www.bluenile.com/education/diamonds) website, carat is the second most important of the 4Cs of diamond. Carat weight choice is a matter of preference, one weight is not necessarily better than another. Typically, heavier diamonds are rarer than ones with lower carat weights and diamond prices can reflect this.

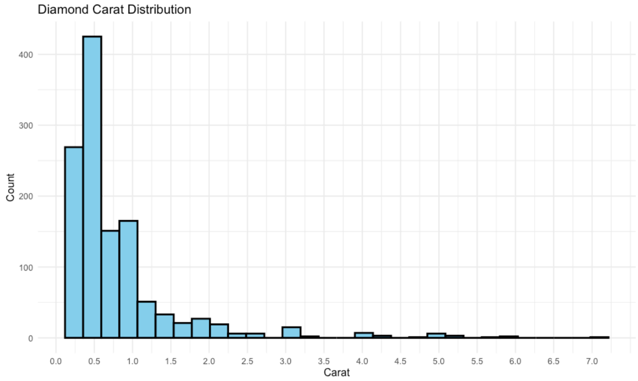


Figure 1.1: Diamond Carat Distribution

Looking at Figure 1.1, the distribution of Diamond Carat it appears with the majority with counts ranging between 0.2-1.0 carats. There is also a sharp decline in frequency as carat size increases.

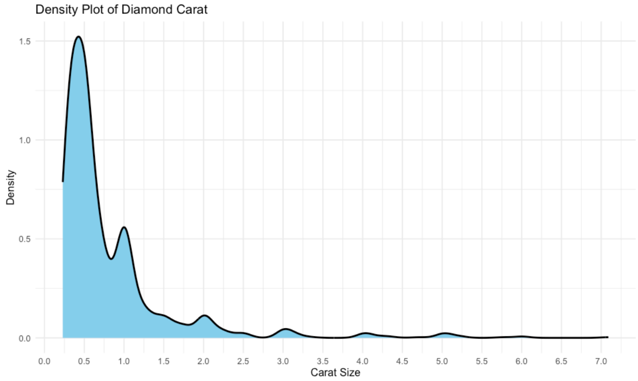


Figure 1.2: Density plot for Carat

Looking at Figure 1.2, The density plot further reinforces the right-skewed distribution observed in the histogram, showing that the majority of diamonds have a small carat size, with the highest density occurring between 0.2 and 1 carats

Table 1.1: Table 1.2: Summary of Diamonds by Carat Range

| Carat Range | Count | Proportion |
|---|---:|---:|
| [0,0.25) | 22 | 0.018 |
| [0.25,0.5) | 397 | 0.327 |
| [0.5,0.75) | 394 | 0.325 |

| | | |
|---|---|---|
| [0.75,1) | 67 | 0.055 |
| [1,1.5) | 198 | 0.163 |
| [1.5,2) | 48 | 0.040 |
| [2,3) | 47 | 0.039 |
| [3,4) | 17 | 0.014 |
| [4,5) | 11 | 0.009 |
| [5,6) | 10 | 0.008 |
| [6,7) | 2 | 0.002 |
| [7,8) | 1 | 0.001 |

Looking at Table **??**, The majority of diamonds fall within the 0.25 to 0.75 carat range, which collectively accounts for about 65.3% of the dataset. The proportion drops significantly beyond 1 carat, with diamonds between 1 and 1 carats making up only 16.3%, and as the carat increases the number of diamonds in those ranges decreases.

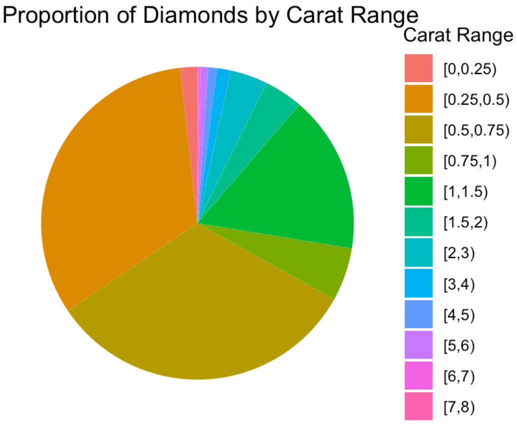


Figure 1.3: Pie graph summarizing Carat proportion

Figure 1.3, The Pie graph visualizes Table **??** and also shows that most diamonds fall udner the range of 0.25-0.5 through 0.5-0.75 carat range. It also shows that there are less of higher carat diamonds which reflects the claim that the heavier diamonds are more rare

# 2 Bivariate Analysis

## 2.1 Bivariate Analysis on Prices

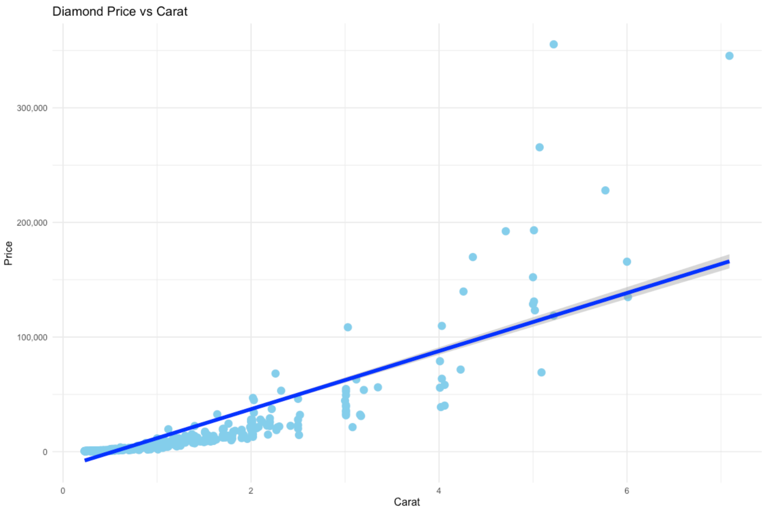


Figure 2.1: Diamond Price vs Carat Size

Figure 2.1 shows the relationship between carat and price with a fitted regression line as the trend. There is a positive correlation between carat and price, meaning that the diamonds with higher carats tend to be more expensive. The regression line shows the overall trend which reflects an increase in price as carat size increases. There are some outliers that appear in the higher carat range which could be due to cut, color, and clarity.

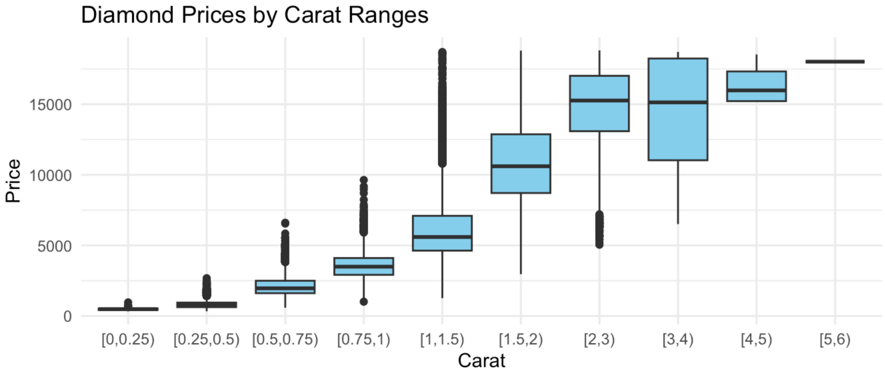


Figure 2.2: Box and whisker plot

The figure 2.2 shows that diamond prices generally increase with carat. The variability in carat rangee and prices for diamonds in the [0, 0.25] carat range are low, while those in the [5, 6] carat range are high. There are outliers in most carat ranges, especially in the mid-range carats, indicating some diamonds are priced significantly higher than others of similar size probably due to the other 3Cs of the 4Cs.

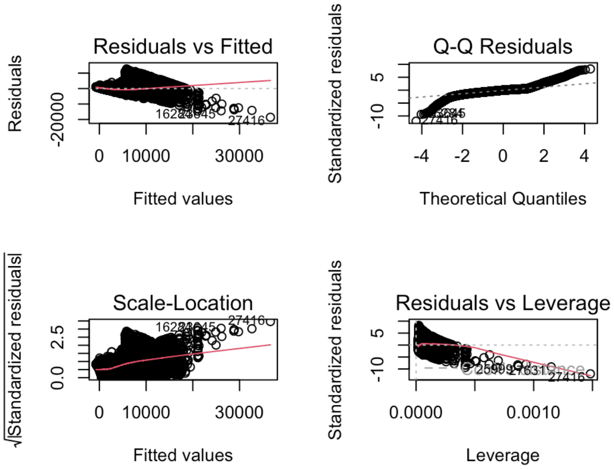
Table 2.1: Table 2.2: Summary of Diamond Prices ($) by Carat Range

| Carat Range | Min Price | Max Price | Average Price | Variability (%) |
|---|---:|---:|---:|---:|
| [0,0.25) | 326 | 963 | 489.56 | 195.40 |
| [0.25,0.5) | 334 | 2677 | 802.55 | 701.50 |
| [0.5,0.75) | 584 | 6607 | 2082.09 | 1031.34 |
| [0.75,1) | 1013 | 9636 | 3550.76 | 851.23 |
| [1,1.5) | 1262 | 18700 | 6139.89 | 1381.77 |
| [1.5,2) | 2964 | 18806 | 10897.17 | 534.48 |
| [2,3) | 5051 | 18823 | 14846.95 | 272.66 |
| [3,4) | 6512 | 18710 | 14308.71 | 187.32 |
| [4,5) | 15223 | 18531 | 16458.00 | 21.73 |
| [5,6) | 18018 | 18018 | 18018.00 | 0.00 |

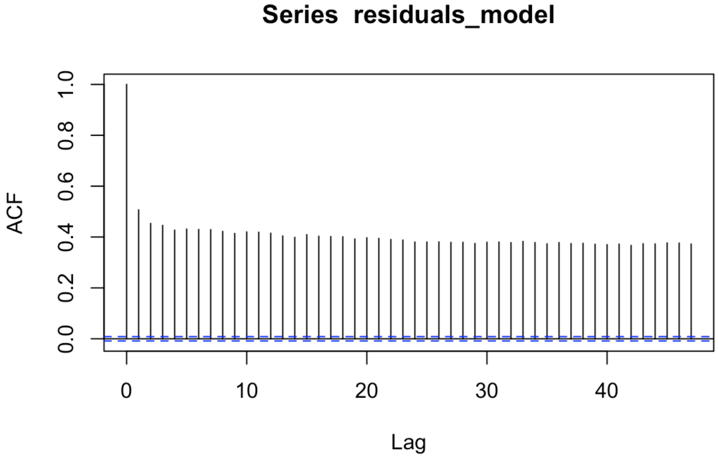
Looking at Table 2.1, there seems to be clear trend of the average and max price increasing as the carat size increases. Smaller diamonds in range 0-0.25 carats have lower average prices and have higher variability, showing price fluctuations which could be influenced by factors such as cut, color, and clarity which also applies to the middle ranges 0.25-1.5 carats. However, in the larger carat ranges 2-6 carats, as the average prices continue to rise, the variability decreases significantly, which tells us that larger diamonds tend to have more predictable pricing, likely due to their rarity of higher carat diamonds.

or section 3, you will be graded on these elements: Describing any transformation performed on the variables when fitting the SLR model, including reasons why these specific transformations were used. Checking SLR assumptions. Providing contextual comments on how the SLR model inform us how price of diamonds are related to carat.

## 2.2 Carat and Price SLR



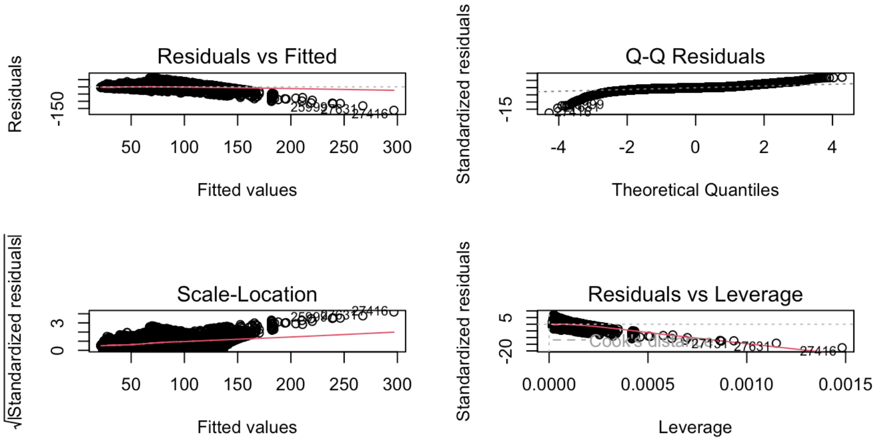
(#fig:Residual_Plot)Residuals plot



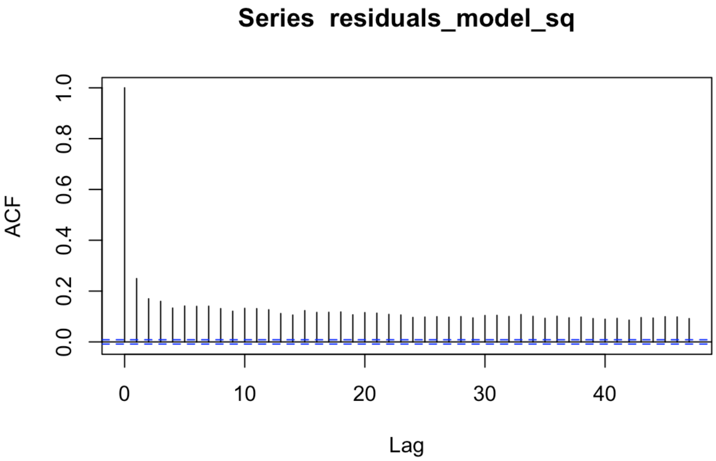
(#fig:ACF_plot)ACF Residual Model

According to Figure @ref(fig:Residual_Plot) the `price` and `carat` models assumptions are:

- `Assumption 1 is not met` : The average value of the residuals for differing values along the x-axis having a apparent curvature in the residual plot.
- `Assumption 2 is not met` : The vertical spread of the residuals not being constant as we move from left to right in the resdiual plot
- `Assumption 3 is not met` : There being High Autocorrelation at Lag 1 as seen from Figure @ref(fig:ACF_plot) which suggests a high dependence in the residuals.
- `Assumption 4 is  met` : The plots do fall closely to the QQ line, which is evidence that the observations follow a normal distribution.



(#fig:Residual_Plot_Trans)Residuals plot



(#fig:ACF_plot_tran)ACF Residual Model

The assumptions having not been met required a transformation. A Square Root transformation was conducted on Price variable to correct the non linear relationship between price and carat.

The transformed models assumption based on figure @ref(fig:Residual_Plot_Trans) with keeping in mind that with real data, assumptions are rarely met 100%.

- `Assumption 1 is met` : The average value of the residuals for differing values along the x-axis does not have an apparent curvature in the residual plot.
- `Assumption 2 is met` : The vertical spread of the residuals is being constant as we move from left to right in the resdiual plot
- `Assumption 3 is met` : There being low Autocorrelation at Lag 1 as seen from Figure @ref(fig:Residual_Plot_Trans) which suggests a low dependence in the residuals.
- `Assumption 4 is  met` : The plots do fall closely to the QQ line, which is evidence that the observations follow a normal distribution.

```
##
## Call:
## lm(formula = price_sqrt ~ carat, data = diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.483   -4.546   -0.962    3.002   66.535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.14643    0.07791   130.2   <2e-16 ***
## carat       57.19923    0.08394   681.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.241 on 53938 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8959
## F-statistic: 4.643e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

The Simple Linear regression model gives an equation of
`Sqrt(Price) = 10.14643 + (57.19923*Carat) - This means that for each 1 unit increase in carat the square root of price increases by 57.1`
value < 2.2e-16`

- This means carat is a strong predictor of price which supports the claim from blue nile that carat has the biggets impact on price. The model gives a `R2 value of 0.8959`
- This tells us that 89.59% of the variation in sqrt(price) is explained by carat.

The regression model effectively shows the relationship between carat and price and showing a strong fit. The model results confirm that carat is a significant predictor of price which means that as the carat increases the price follows a trend that is predictable.