# STA302 Final Project Part 1: Project Proposal

[Research Question]

Yolanda Thant, Kosar Hemmati, Shameiza Hussain, Ayushi Verma

15 May 2025

# Introduction

When considering a diamond's value, many consumers rely on a grading system introduced by the Gemological Institute of America (n.d.) called the "4Cs": carat, cut, clarity, and colour. This system was designed to help consumers of the diamond market understand the value of their purchases based on the values assigned to each of these four traits. While information about factors that influence diamond pricing is publicly available, the numeric impact of each factor is unclear (Lee et al., 2014; Mamonov & Triantoro, 2018). For instance, two diamonds of the same weight can differ vastly in price after accounting for cut, clarity, and colour (Tiffany & Co., n.d.), but the explanation behind this variance has not been rigorously quantified based on our literature research. Thus, our research addresses the question: To what extent do carat, colour, clarity, table size, and length predict the price of a diamond? While previous studies have conducted similar analyses (Mamonov & Triantoro, 2018; Özmen, 2024), our goal is to later refine our model in hopes of explaining more of the observed price variation in our data. Quantifying the effects that different features have on diamond prices can better inform consumers' judgments on whether they are paying a fair price.

Multiple linear regression (MLR) is a suitable method to answer our question, because we are exploring the association between price, a continuous response variable, with several explanatory variables simultaneously. This way, we can infer which diamond factors are the strongest predictors of price. MLR models the (estimated) average of the response versus given values of the predictors, which is relevant to our goal of helping consumers understand where the price of their diamond lies in relation to the estimated average price for their specific diamond profile. The focus of the model will be on interpretability since we are more interested in drawing conclusions about the average effects of our chosen diamond attributes on the price, rather than trying to predict exact prices of all diamonds.

# Exploratory Data Analysis (EDA) - Data Description

The selected dataset was retrieved and downloaded from Kaggle, a data science platform, and was uploaded by a university instructor for their students to practice data analysis skills. There was no explicit research question cited on the Kaggle posting, and our research question was constructed after a careful review of the dataset and the corresponding literature surrounding diamond pricing. Further investigation of the dataset revealed that the data was originally published in the ggplot2 R package. In particular, the diamond dataset was collected and curated by Wickham et al., the creators and contributors of ggplot2, using the Loose Diamond Search Engine, which contains historical data on diamond prices and descriptions of their various physical attributes.

## Response Variable Summary

When determining the market price of diamonds, jewelers justify their valuation with reference to the physical attributes of the diamond such as its cut, carat, colour, and size. Thus, price is an appropriate selection for the response variable as it is dependent on the chosen predictors. Further, price is a continuous numerical value measured in US Dollars, making it suitable for linear regression analysis.

Table 1: Numerical Summary of Response Variable

| Statistic | Value |
|---|---:|
| Minimum | 326.000 |
| 1st Quartile | 950.000 |
| Median | 2401.000 |
| Mode | 605.000 |
| Mean | 3933.000 |
| Standard Deviation | 3989.338 |
| 3rd Quartile | 5324.000 |
| Maximum | 18823.000 |

## Predictor Variable Summaries

Table 2: Numerical Summary of Quantitative Predictor Variables

| Statistic | Carat | Table | x |
|---|---|---|---|
| Minimum | 0.2000 | 43.0000 | 0.0000 |
| 1st Quartile | 0.4000 | 56.0000 | 4.7100 |
| Median | 0.7000 | 57.0000 | 5.7000 |
| Mode | 0.3000 | 56.0000 | 4.3700 |
| Mean | 0.7979 | 57.4573 | 5.7312 |
| Standard Deviation | 0.4740 | 2.2345 | 1.1217 |
| 3rd Quartile | 1.0400 | 59.0000 | 6.5400 |
| Maximum | 5.0100 | 95.0000 | 10.7400 |

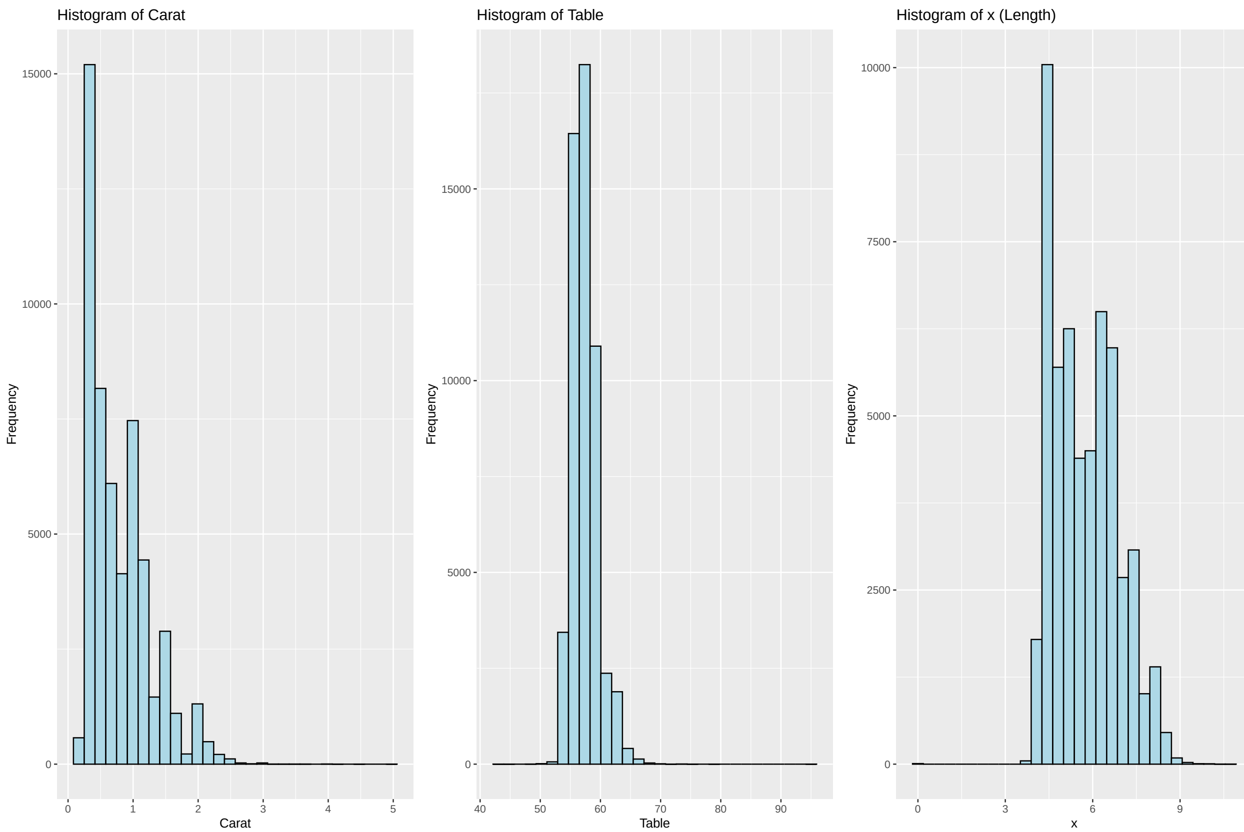
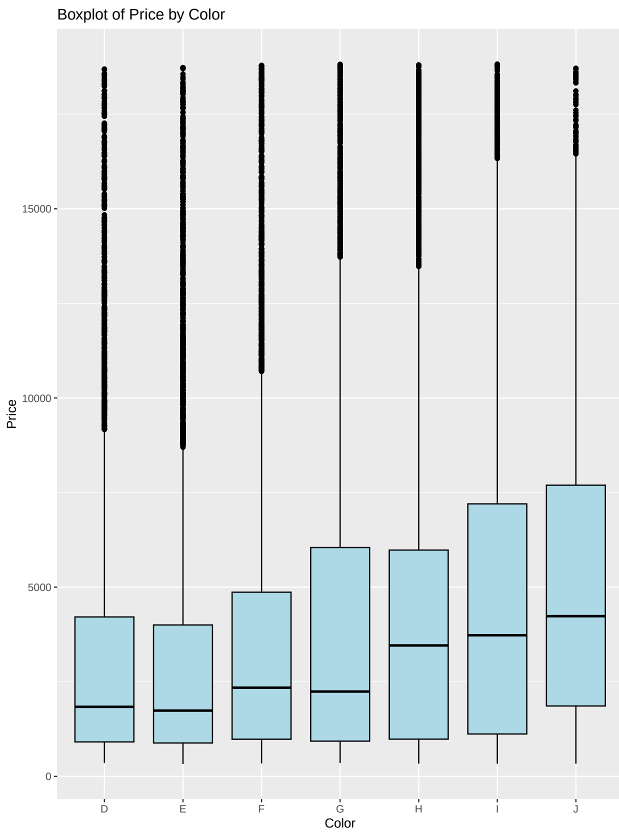


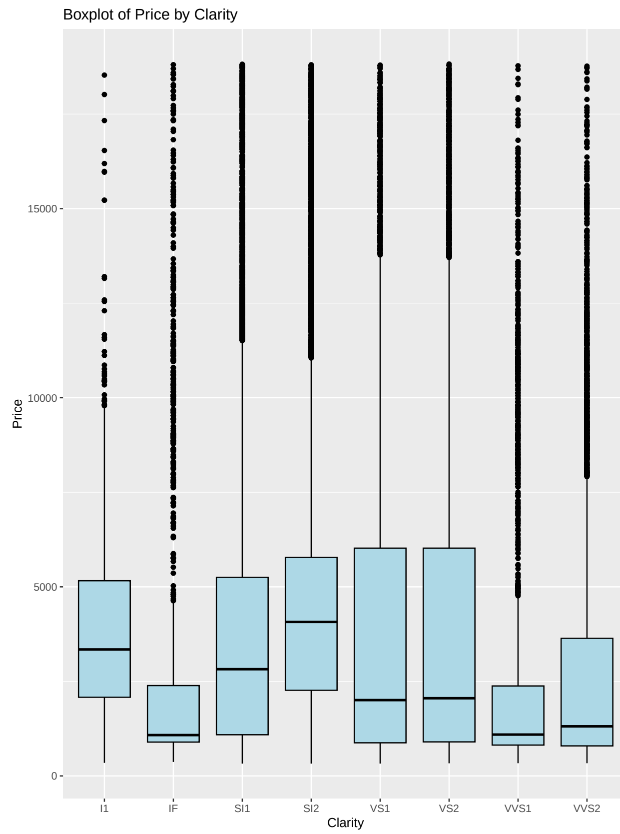

Table 3: Numerical Summary of ‘Color’ Predictor

| Color | Frequency | Proportion |
|---|---|---|
| D | 6775 | 0.1256 |
| E | 9799 | 0.1817 |
| F | 9543 | 0.1769 |
| G | 11292 | 0.2093 |
| H | 8304 | 0.1539 |
| I | 5422 | 0.1005 |
| J | 2808 | 0.0521 |

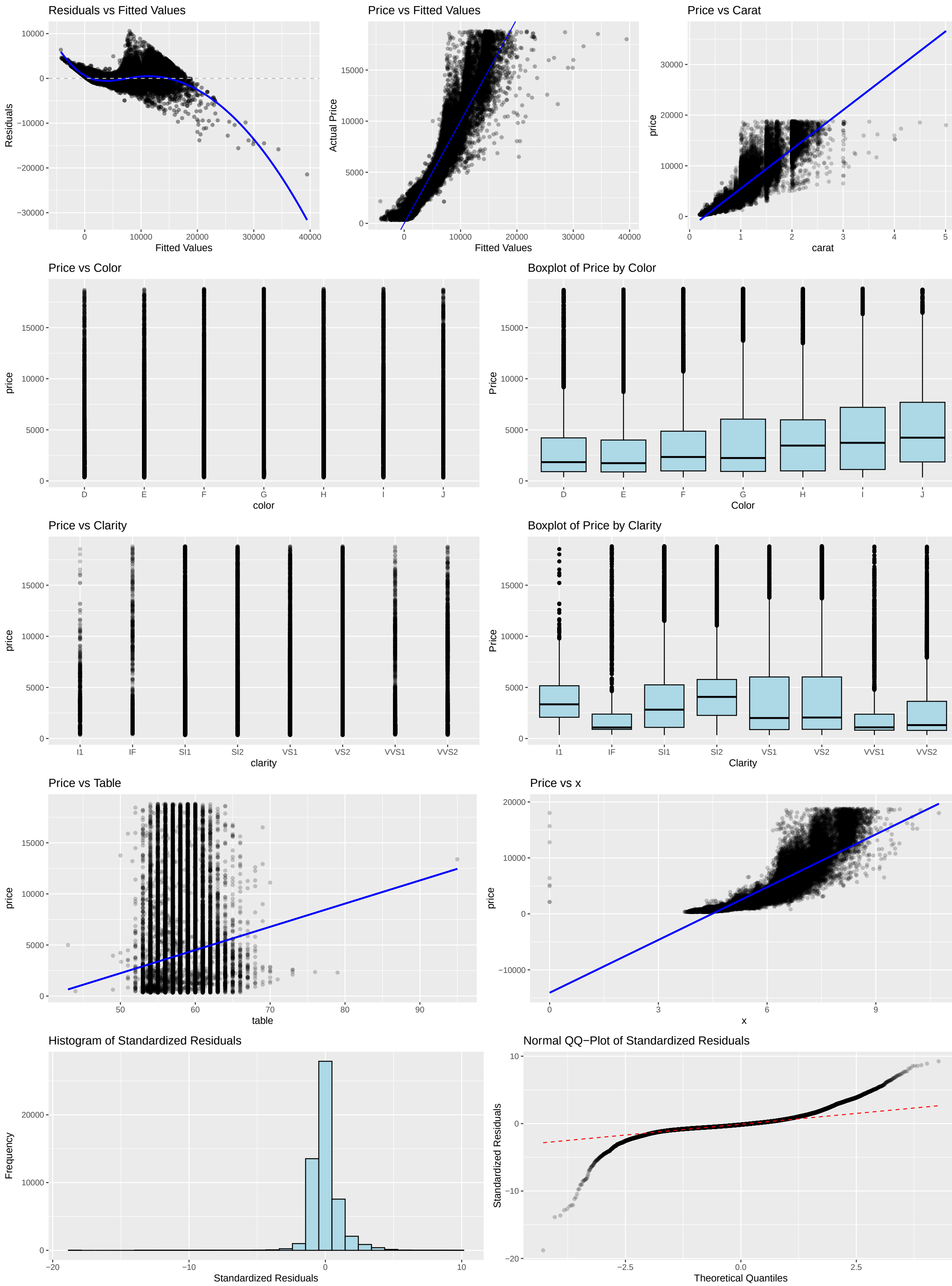
Table 4: Numerical Summary of 'Clarity' Predictor

| Clarity | Frequency | Proportion |
|---|---|---|
| I1 | 741 | 0.0137 |
| IF | 1790 | 0.0332 |
| SI1 | 13067 | 0.2422 |
| SI2 | 9194 | 0.1704 |
| VS1 | 8171 | 0.1515 |
| VS2 | 12259 | 0.2273 |
| VVS1 | 3655 | 0.0678 |
| VVS2 | 5066 | 0.0939 |



# Preliminary Model Results



$$
\begin{aligned}
y = \beta_0 &+ \beta_1 \cdot \text{carat} + \beta_2 \cdot I(\text{color} = E) + \beta_3 \cdot I(\text{color} = F) + \beta_4 \cdot I(\text{color} = G) + \beta_5 \cdot I(\text{color} = H) \\
&+ \beta_6 \cdot I(\text{color} = I) + \beta_7 \cdot I(\text{color} = J) + \beta_8 \cdot I(\text{clarity} = \text{IF}) + \beta_9 \cdot I(\text{clarity} = \text{SI1}) \\
&+ \beta_{10} \cdot I(\text{clarity} = \text{SI2}) + \beta_{11} \cdot I(\text{clarity} = \text{VS1}) + \beta_{12} \cdot I(\text{clarity} = \text{VS2}) \\
&+ \beta_{13} \cdot I(\text{clarity} = \text{VVS1}) + \beta_{14} \cdot I(\text{clarity} = \text{VVS2}) + \beta_{15} \cdot \text{table} + \beta_{16} \cdot x + \epsilon
\end{aligned}
$$

$$
\text{Assume } \mathbb{E}[\epsilon] = 0, \mathbb{V}[\epsilon] = \sigma^2, \text{and } \epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)
$$

$$
\begin{aligned}
\hat{E}[y] = \hat{y} = \ &\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{carat} + \hat{\beta}_2 \cdot I(\text{color} = E) + \hat{\beta}_3 \cdot I(\text{color} = F) + \hat{\beta}_4 \cdot I(\text{color} = G) + \hat{\beta}_5 \cdot I(\text{color} = H) \\
&+ \hat{\beta}_6 \cdot I(\text{color} = I) + \hat{\beta}_7 \cdot I(\text{color} = J) + \hat{\beta}_8 \cdot I(\text{clarity} = \text{IF}) + \hat{\beta}_9 \cdot I(\text{clarity} = \text{SI1}) \\
&+ \hat{\beta}_{10} \cdot I(\text{clarity} = \text{SI2}) + \hat{\beta}_{11} \cdot I(\text{clarity} = \text{VS1}) + \hat{\beta}_{12} \cdot I(\text{clarity} = \text{VS2}) \\
&+ \hat{\beta}_{13} \cdot I(\text{clarity} = \text{VVS1}) + \hat{\beta}_{14} \cdot I(\text{clarity} = \text{VVS2}) + \hat{\beta}_{15} \cdot \text{table} + \hat{\beta}_{16} \cdot x
\end{aligned}
$$

Table 5: Numerical Summary of Coefficient Estimates

| Term | Estimate |
|---|---:|
| (Intercept) | -1388.839 |
| carat | 10945.695 |
| colorE | -210.489 |
| colorF | -286.752 |
| colorG | -493.858 |
| colorH | -993.809 |
| colorI | -1474.570 |
| colorJ | -2387.941 |
| clarityIF | 5665.892 |
| claritySI1 | 3894.266 |
| claritySI2 | 2926.092 |
| clarityVS1 | 4845.059 |
| clarityVS2 | 4528.148 |
| clarityVVS1 | 5309.983 |
| clarityVVS2 | 5234.228 |
| table | -32.708 |
| x | -900.270 |

# Residual Analysis

The residual analysis revealed significant violations of core regression model assumptions: linearity, constant error variance (homoscedasticity), and normality of errors. The **Residuals vs Fitted Values** plot showed a clear, curved LOESS trendline with increasing spread as fitted values increased, suggesting non-linearity and heteroscedasticity.

Similarly, the **Price vs Fitted Values** plot showed a generally strong positive relationship, but with slight curvature and growing variance at higher price levels, indicates non-linearity & heteroscedasticity. The **Price vs Carat** scatterplot suggested an exponential relationship that a smile linear model fails to capture reasonably.

At first glance, categorical variables **color** and **clarity** don't appear to be strong predictors of price due to seemingly constant spread of points across levels. However, the boxplots reveal significant stratified effects on **price**, confirming that these variables do show variation.

For **color**, although the maximum prices are similar across all levels, the median prices differ, with diamonds in **color** J having higher median prices, and D showing higher outliers.

Diamonds with **clarity** I1 show outliers and higher median prices, whereas those with IF appear to have more outliers and lower medians. This stratification reinforces the idea that both are valid categorical predictors for diamond price modeling.

However, the **Price vs Table** plot showed a weak and noisy association, with data points highly concentrated in a narrow range, suggesting that the table predictor may not be a strong predictor to predict price. The **Price vs x** plot also showed curvature, suggesting non-linearity.

Furthermore, the **histogram of standardized residuals** sharply peaked around 0 with heavy tails, and the **normal QQ-plot** confirmed it as it showed deviations at both tails, indicating violations of the normality assumption.

# Model Discussion

From model estimates, a diamond with **color** D, **clarity** I1, and 0 values for **table**, **x**, and **carat**, is predicted to have an average price of -$1,388.389. Though this scenario isn't practically meaningful, it serves as the model's (baseline) reference level. When the diamond's **color** is E and **clarity** is IF, with the same 0 values for the continuous variables, the (estimated) average price increases to $4,066.564, reflecting the strong positive contribution of high **clarity** IF to the overall **price**.

Moreover, a 1-gram increase in **carat**—with the same values as the previous example—raises the predicted average price by $10,945.695 per gram of **carat**, confirming that **carat** is the most influential predictor.

Interestingly, **table** and **x** are both negatively associated with **price**, with coefficients of −$32.71 and −$900.27, respectively. This is counterintuitive as larger dimensions are generally more desirable. **x** likely reflects multicollinearity with **carat**, since both capture size-related information. This redundancy suggests the need to reconsider variable inclusion—possibility removing **x** or replacing with a composite variable like volume.

# Bibliography


Al Aswad, Ms. N. (2022, July 9). *Diamonds Prices.* Kaggle. https://www.kaggle.com/datasets/nancyalaswad90/diamonds-prices

Diamond Prices Comparison. Loose Diamonds Search Engine. (n.d.). https://www.diamondse.info/diamond-prices.asp

Gemological Institute of America. (n.d.). *4Cs of diamond quality.* Retrieved May 20, 2025, from https://4cs.gia.edu/en-us/4cs-of-diamond-quality/

Lee, J. A., Caudill, S. B., & Mixon, F. G. (2014). Shine bright like a diamond: A hedonic model of grading and pricing an experience good. *Applied Economics, 46*(16), 1829–1838. https://doi.org/10.1080/00036846.2014.884707

Mamonov, S., & Triantoro, T. (2018). Subjectivity of diamond prices in online retail: Insights from a data mining study. *Journal of Theoretical and Applied Electronic Commerce Research, 13*(2), 15–28. https://doi.org/10.4067/s0718-18762018000200103

Özmen, G. (2024, August 12). *More than carat: A comprehensive regression analysis for diamond price prediction.* Medium. https://medium.com/@gizemzmen/more-than-carat-a-comprehensive-regression-analysis-for-diamond-price-prediction-b9222f337154

Tiffany & Co. (n.d.). *What Is A Diamond Carat?* Retrieved May 20, 2025, from https://www.tiffany.ca/engagement/the-tiffany-guide-to-diamonds/carat/

Wickham, H., van den Brand, T., Dunnington, D., Yutani, H., Woo, K., Wilke, C., Takahashi, K., Pedersen, T. L., Henry, L., & Chang, W. (n.d.). *Prices of Over 50,000 Round Cut Diamonds*. ggplot2. https://ggplot2.tidyverse.org/reference/diamonds.html#ref-usage