Diamond Price Analysis

Samantha Howard

CMSE 492

4/01/2022

Table of Contents

List of Figures

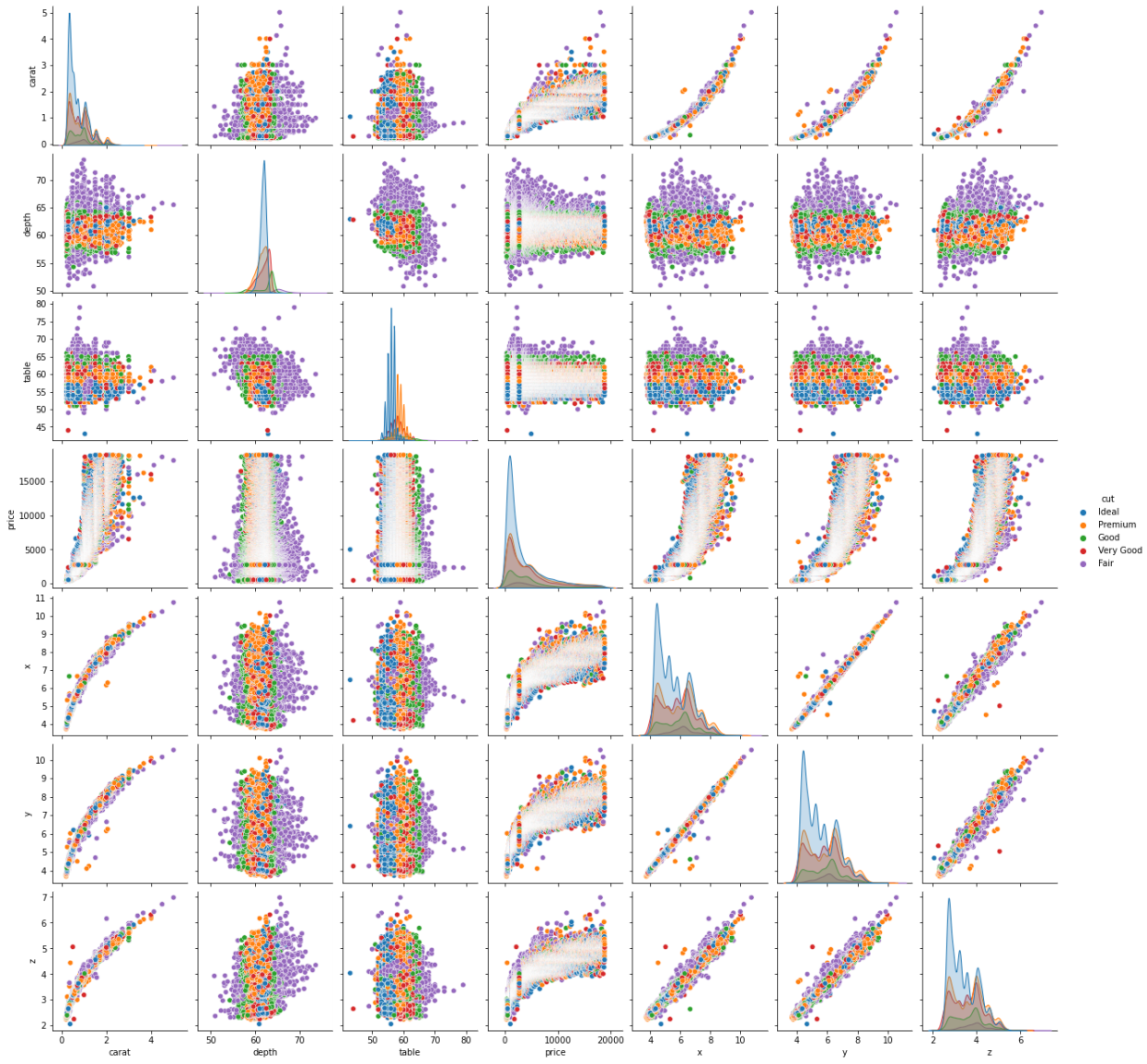

Figure 1: Seaborn Pairplot of the variables, color is based on cut

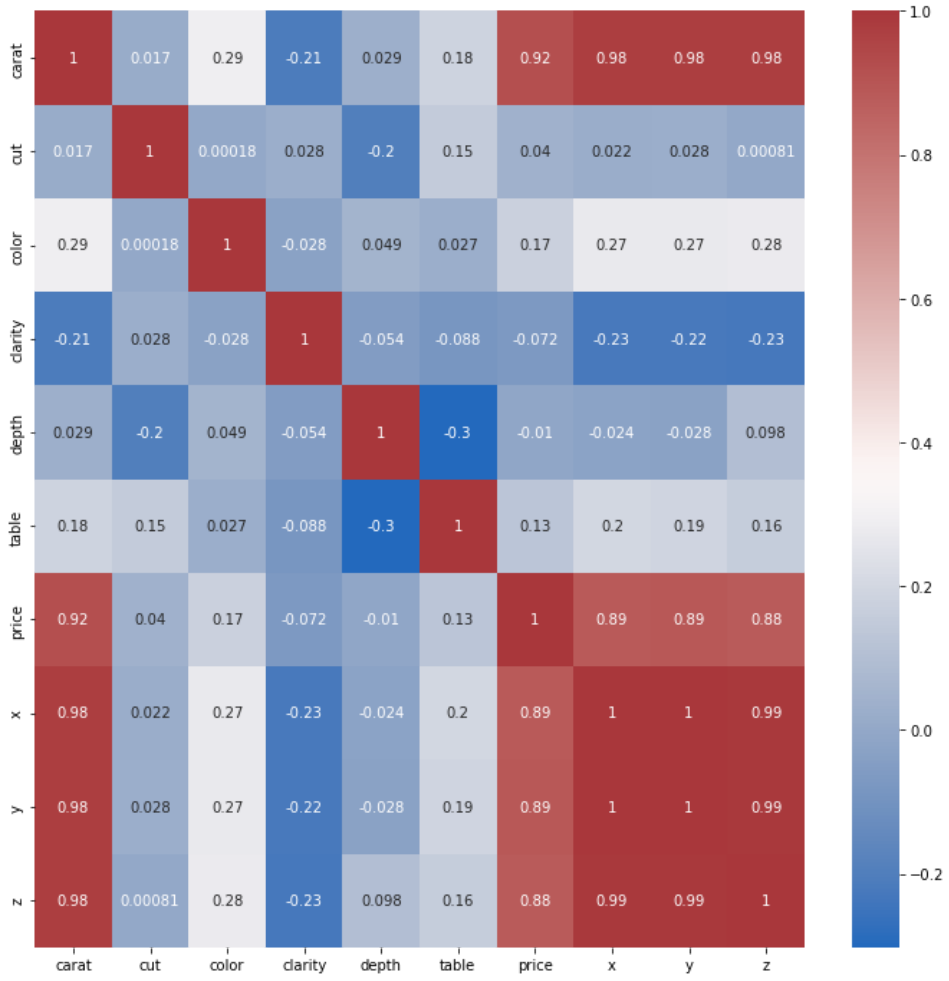

Figure 2: Seaborn Heatmap of the correlation

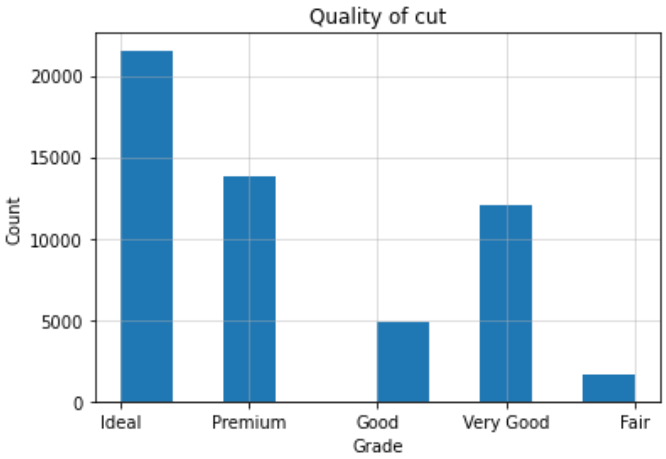

Figure 3: Histogram of Quality of the Cut

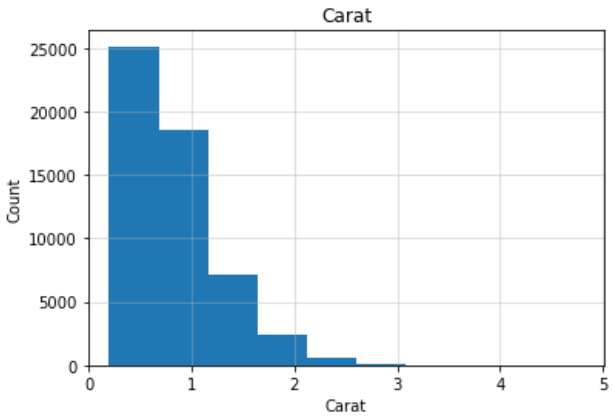

Figure 4: Histogram of Carrot Distribution

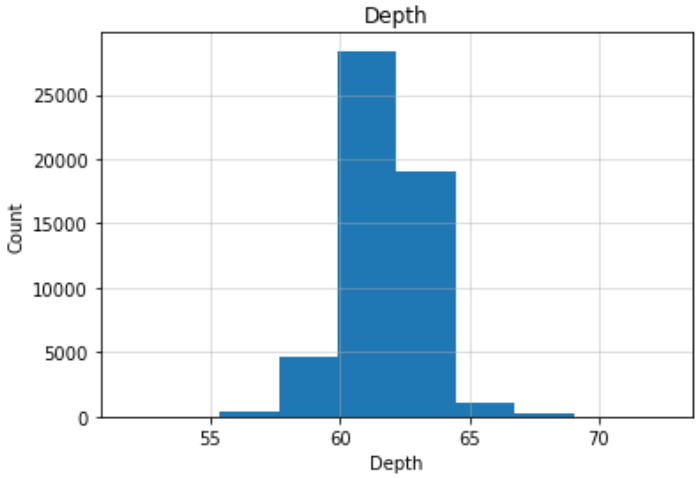

Figure 5: Histogram of Depth Distribution

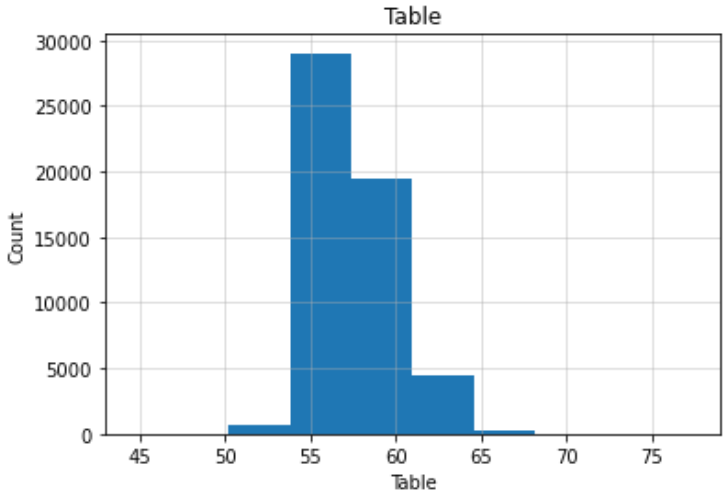

Figure 6: Distribution of the Variable “Table”

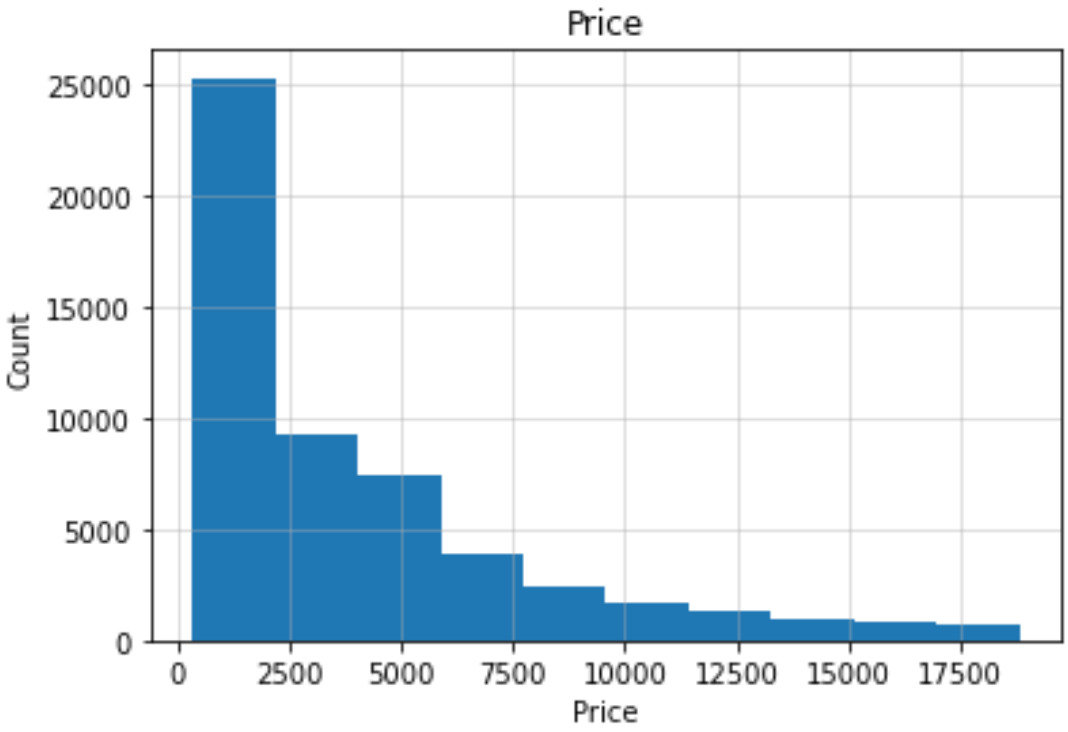

Figure 7: Distribution of Price

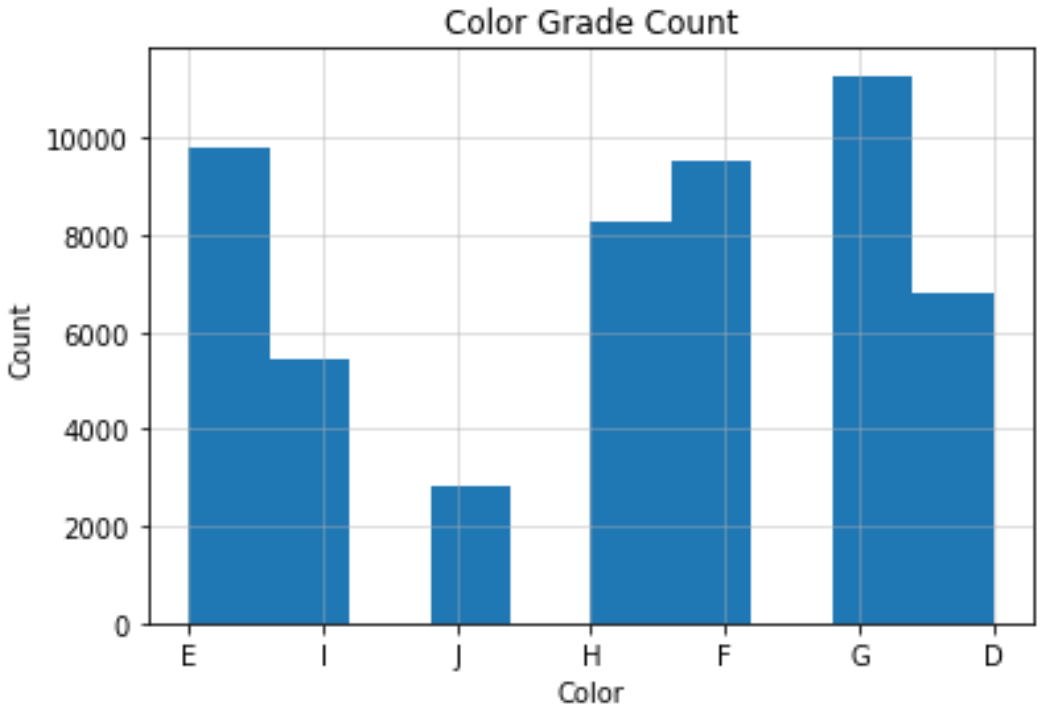

Figure 8: Color Distribution

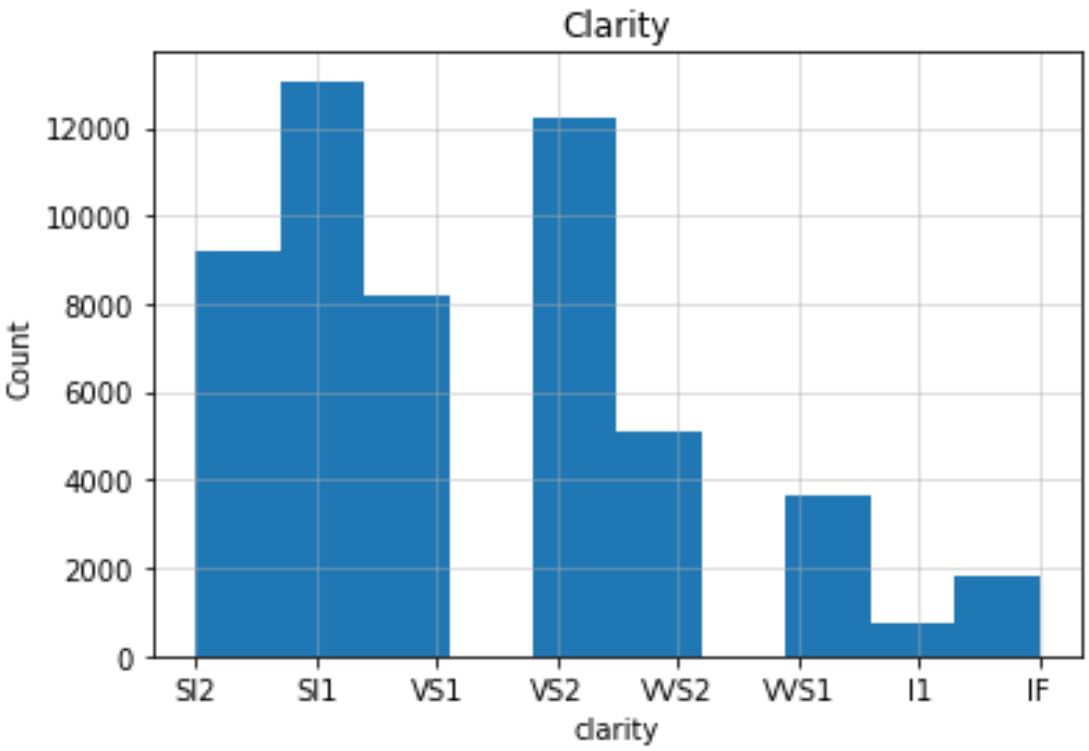

Figure 9: Histogram of Clarity

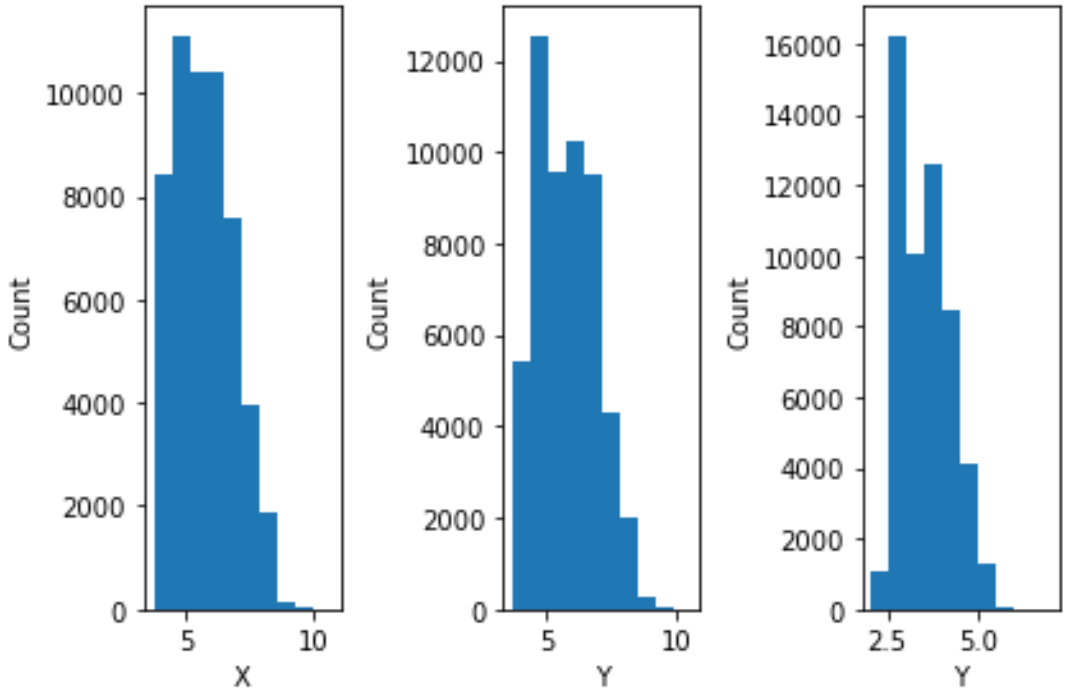

Figure 10: Subplot Histograms of Sizes

Introduction

For my CMSE 492 Project I am using a large dataset from Kaggle to predict diamond prices. I chose this project as a homage to when I originally encountered this data set back in STT 180 as an in class assignment using R. Coming back to it a few years later, and in a different language, I thought it could be well suited as a final project for this class.

For this project I will be using multiple machine learning algorithms to see what one predicts cost the most effectively given the other features. Currently I am using five different machine learning methods from various python packages to determine what one worked best. I plan to further develop these parameters by adjusting parameters or using other methods to increase the accuracy.

The final deliverable of this project will be a written report, a project poster, and a python notebook. The written report will be similar to this, with explanations of the code, methods and including finalized figures. A poster that will summarize the information neatly. A python notebook, containing all the code and generated figures.

Methodologies

As of right now I have used five different machine learning methods; LinearRegression, DecisionTree, RandomForest, KNeighbors, XGBRegressor. The one that performed the best of these five was XGBRegressor. The R-squared value was calculated to be 0.9811 and the adjusted R-squared was calculated to be similar, being the same when rounded to four significant figures. These results were obtained by predominantly using 5 different packages, and many imports, mostly sklearn, pandas, numpy, matplotlib and xgboost. For the machine learning algorithm that performed the best, xgboost, I calculated: R^2, adjusted R^2, MAE, MSE, and RMSE values.

- Task Definition: formally define what you are doing (inputs, outputs)
- 3. Algorithm Definition: what algorithms did you use? (e.g., RL or DR or classification)
- 4.Machine Learning Approach (see items above)
    - A. DS/EDA methods
    - B. ML approach/pipelines/tuning
    - C. results, visualizations, comparisons
    - D. discussion of results and conclusions about your go

Conclusion

In conclusion I have the bulk of the code completed, with the intention of polishing the figures and seeing if the accuracy of the machine learning aspect can be further increased. For this data set I have the R-squared, adjusted R-squared, MAE, MSE and RMSE currently calculated for the machine learning algorithm that performed the best. I consider these current results to be exelect but can be further polished in the time remaining before the deadline. In terms of the visualizations I wish to change the correlation plot to something more easily digestible and consider changing the histograms to violin plots or another plot type that could convey a little more information. In addition to these things my next primary goal is to add more explanatory writing to the notebook to express what is going on at various steps of the project.

However, I'm currently unaware of documentation outlying other expectations to this project other than this current assignment that requested a write up of the progress; with that being said I hope what I have outlined is satisfactory as I am unsure of my expectations for a final deliverable for this project.

- recap your goal and what was achieved
- recap the methods you used
- what did you conclude from your study?
- future work: what were the shortcomings, what improvements can you propose?

References

Data Allocation:

Data: from Kaggle

Generalized Documentation for the Packages Used:

Numpy:

Pandas:

Seaborn:

MatPlotLib:

SKLearn:

XGBoost:

Stack Overflow:

Histogram Range:

Miscellaneous:

Proposal Writing:

How to calculate a Diamond's weight in Carats. (n.d.). Retrieved April 14, 2022, from https://www.jewelrynotes.com/how-to-calculate-a-diamonds-weight-in-carats/#:~:text=To%20calculate%20the%20carats%20of,stone%20weighs%20half%20a%20carat.