Advanced Machine Learning – Final Project Write-up
Project: Clustering Analysis of Diamond Data for Quality Assessment

By: Sarah Hazziza (337891675) & Niv Levi (314628165)
Github Repository: https://github.com/nivlevi1/ML2_Final_Assignment

Abstract

While the traditional 4Cs – Cut, Color, Clarity, and Carat Weight – are widely recognized as key determinants of a diamond's quality and value, this study aims to investigate the true strength and importance of these factors in defining distinct clusters or groups of diamond. Leveraging a rich dataset, crawled from the web, with Gemological Institute of America (GIA) and International Gemological Institute (IGI) diamond grading characteristics, we analyzed the dataset thanks to data preprocessing, Exploratory Data Analysis (EDA), and Principal Component Analysis (PCA) to effectively reduce dimensionality while preserving maximal variance. As expected, we revealed the 4Cs' prominence across the first principal components. To discern natural groupings within the dataset, three distinct unsupervised clustering algorithms – K-means, DBSCAN, and Agglomerative clustering – were applied. After evaluation of these algorithms using silhouette score, sum of squared error as internal validation metrics and of course visualization, the k-means algorithm yielded cohesive and well-separated clusters, leading us to use its results to understand the different characteristics of each cluster. This project builds upon previous research focused on price prediction by addressing the need for quality assessment in diamond evaluation. By leveraging clustering techniques, we aim to provide a more comprehensive understanding of diamond characteristics and their impact on market value, thereby aiding both buyers and sellers in making informed decisions.

Introduction

Diamonds have long been admired for their beauty and symbolism, with buyers typically focusing their attention on well-known factors like the 4Cs – Cut, Color, Clarity, and Carat. However, beneath this surface lies a realm of lesser-known parameters that wield significant influence over a diamond's price and appearance.

Take fluorescence, for example, this often-overlooked characteristic refers to a diamond's tendency to emit a soft glow when exposed to ultraviolet light. While fluorescence can enhance the appearance of some diamonds, giving them a mesmerizing luminescence, excessive fluorescence may impart a hazy or milky appearance, detracting from their overall beauty and value.

Furthermore, the diamond industry encompasses both natural and lab-grown diamonds, each with its own distinct characteristics and market dynamics. The Gemological Institute of America (GIA) and the International Gemological Institute (IGI) play pivotal roles in grading and certifying diamonds, providing consumers with valuable information about a diamond's quality and provenance. But, of course, for consumers to make the most of this information, they need to understand what it means. Diamond sellers, capitalizing on this knowledge gap, may omit details about certain characteristics or prioritize marketing based solely on the 4Cs, leading to inflated prices and limited choices for consumers.

To address this challenge, our project employs advanced machine learning techniques to uncover hidden patterns and relationships within diamond data. By applying clustering algorithms to a comprehensive dataset, we aim to discern natural groupings of diamonds based on a wide range of parameters. Through this analysis, we seek to provide buyers with a deeper understanding of diamond quality and value, empowering them to make informed decisions that align with their preferences and priorities.

Ultimately, our endeavor is to foster transparency and accountability within the diamond industry, ensuring that buyers are equipped to navigate the market confidently and confidently discern true value from superficial marketing tactics.

Dataset and Features

For our project, we employed web crawling techniques to build our diamond dataset. The website whiteflash is a dynamic website which implied from us learning new techniques of crawling using the Selenium library, a powerful tool for browser automation. The crawling procedure involved several key steps to systematically retrieve and structure the desired data. On the website, the natural and the lab diamonds are separated explaining why we performed 2 web crawling, then took care of the structure of the 2 dataset retrieved to combine them effectively.

The first step was to set up the webdriver, Selenium uses a webdriver functionality to automate web interactions. We used the Edge WebDriver enabling programmatic control over the browser. This technique allows navigating through web pages and extract relevant content. After setting up the webdriver, we had to understand the website's URL where the data of interest is hosted to allow our crawler to access the relevant website for data collection. We built a loop to iterate through pages, and programmed the crawler to scroll down each webpage triggering the loading of dynamically appearing elements and incorporated waiting mechanisms to ensure the full loading of those elements before the data extraction. After trying different methods, we understood this one was necessary since the dynamic loading mechanism was not triggered by simply accessing the new URL with the page number. Still employing Selenium we extracted the HTML content and used BeautifulSoup library to parse the content and navigate through the document structure extracting specific data elements. This step involved identifying and isolating relevant information such as prices, cuts and others to build the basis of our database. As said before, we performed this on both Natural diamonds URL and Lab diamonds URL then concatenated the two Dataframes into one along the rows before storing the dataset in a CSV file for easier access in a new notebook dedicated to our preprocessing and model steps.

After performing this crawling procedure, we moved forward to the preprocessing steps of the data before continuing with our experiments. Firstly, we extracted further information from the relatively raw data scraped, such as colors, clarity and carats of the diamonds. We took care of columns with missing values that could be rectified thanks to informations from other columns, we created new columns length, width and depth thanks to the measurements column, we decided to let go of some data such as diamonds from the AGS certification judging they were too little and that could impact our overall model and motivations. We let go of some irrelevant columns such as links or the ones related to the buying process of the website. To take care of missing values in the column girdle, we decided to fill them with the most present value from the column which was 'medium' for 2 reasons, the first being that the medium is usual for the girdle of diamonds and also since by checking multiple certificate of diamonds from the website, we actually understood it was often medium. Then with this column, we chose to divide it into: thinnest and thickest. The girdle of a diamond usually has 2 measures because of the facets added to it, thanks to knowledge from the web about diamonds we thought that dividing it into 2 columns will be the most appropriate. One of the other preprocessing steps was letting go of columns that showed too much zero values (e.g. Star%, Crown%...) and for the ones having not too many zeros we decided to remove the diamonds showing some zeros in those columns (e.g. Crown Angle and Pavilion Angle).

After those preprocessing steps, we were let with those diamond's grading parameters (columns):

Carat Weight: Larger diamonds are rarer and more valuable. Higher carat weight increases the price significantly.

Cut: The cut has a major effect on a diamond's brilliance and fire. An excellent cut reflects light better, making the diamond more brilliant and valuable. A poor cut reduces brilliance and lowers the price.

Color: Less color is preferred. Colorless (D-F) diamonds are the most valuable. More tinted colors (K-Z) reduce the price.

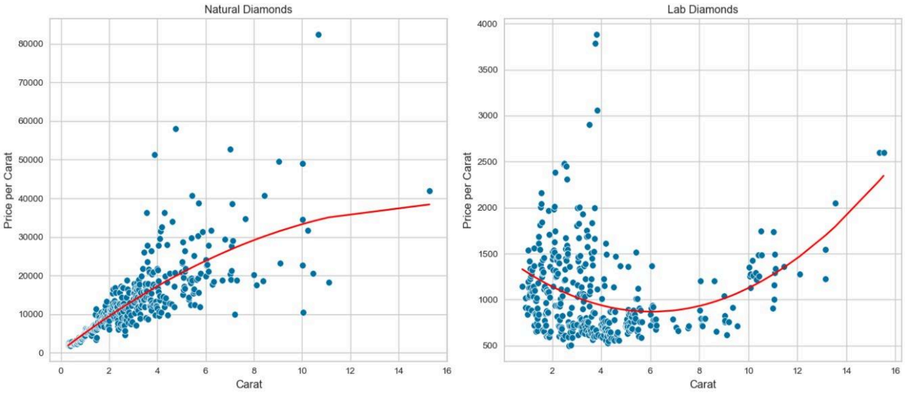
Clarity: Fewer inclusions and blemishes are better. Flawless diamonds are very rare and expensive. Included grades with visible flaws lower the brilliance and price.

Fluorescence: Fluorescence refers to a diamond's reaction to ultraviolet light, where it may emit a soft glow. While fluorescence is not always considered a negative trait, diamonds with strong

fluorescence may appear milky or hazy in certain lighting conditions, affecting their appearance and potentially lowering their value.
Symmetry and Polish: While often considered less important than the 4Cs, the symmetry and polish of a diamond can greatly influence its brilliance and overall visual appeal. Diamonds with excellent symmetry and polish reflect light more effectively, resulting in a more dazzling appearance.
L/W%: The ideal L/W ratio range of 1.00 to 1.03 ensures optimal light performance, resulting in maximum brilliance and fire. A diamond with an L/W ratio within this range will appear visually balanced and appealing, neither too elongated nor too squat. Diamonds with an ideal L/W ratio are rarer and more valuable due to the skilled cutting required to achieve precise proportions.
Depth%: The depth percentage measures the height of a diamond relative to its width. The ideal depth percentage for a round diamond is 59-63.5% for optimal light performance. Too deep or too shallow impacts brilliance and light leakage. It's a key proportion to consider along with table and crown angles.
Table%: The depth and table percentage of a diamond (the height of the diamond relative to its diameter and the size of the table facet, respectively) can impact its brightness and fire. Diamonds with proportions outside of ideal ranges may appear dull or lackluster.
Pavilion Angle: The angle between the bottom facets and the girdle. An ideal pavilion angle (around 40.6°) maximizes light return through the top.
Crown Angle: The angle between the top facets and the girdle. An ideal crown angle (around 34.5°) ensures light enters and refracts evenly. Ideal angles optimize brilliance and value. Non-ideal angles reduce brilliance and price.

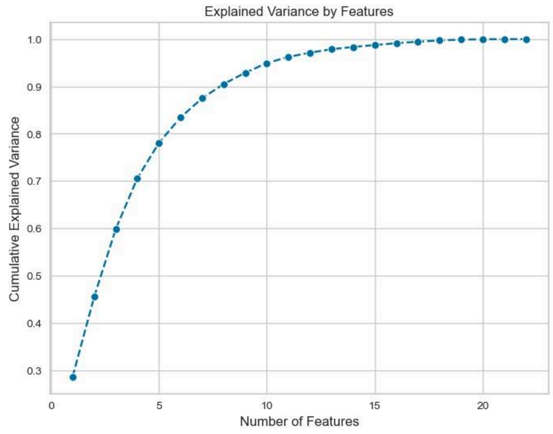
We moved on to the Exploratory Data Analysis Step (EDA) and focused on understanding the characteristics of the diamonds to further understand the problem as well as our dataset. We first tried to understand collinearity between columns,seeing in the heatmap that Length, Width, Depth and Carat were so related to each other, we seek advice and decided to let go of Depth column and perform feature engineering to get the ratio Length/Width%. Then we plotted distribution across the columns thanks to pieplot for categorical columns and histograms for the numerical ones. Overall the data seemed to be well distributed and constituted of various data, implying potential good results for clustering. Some outliers remained but we hoped they could be informative for future steps.
We also performed an EDA differentiating between Natural and Lab diamonds to understand where some differences could show up in future steps. Even though our dataset is constituted of only 12% of Lab diamonds, we judged those nuances could be relevant for our model. We showed that the price of Natural and Lab diamonds behave differently thanks to polynomial functions (see Graph 1). For Natural diamonds, bigger the Carat, higher the price and for Lab diamonds smaller carats are expensive then we see a decrease around carat of size 6.0 then increasing again. Those results may indicate that the price of Lab diamonds is based on demand (high demand for small carats and high demand for high carats because of less Natural diamonds with high carats) and the price of Natural diamonds is based on rarity. Furthermore, we plotted a graph for each categorical column of: category in function of Price per Carat, this one was very informative and we were able to identify for most of the columns the order of the characteristics that are ordinal in diamonds, such as Color or Cut. Indeed we saw an increase of average price per carat when getting to a better level of color or cut of diamond, for Lab and Natural diamonds.

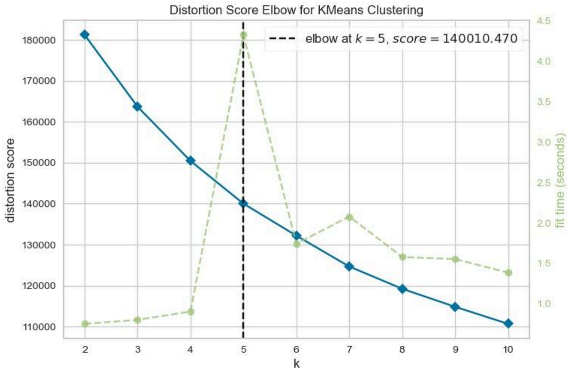
Graph 1:



Methodology

After the preprocessing steps and the EDA, having gathered enough knowledge of our dataset, we started to build our model, especially taking care of the dataset to be able to use clustering algorithms but also trying different algorithms to find the best suited one for our problem. We first created a pipeline to transform our dataset: on ordinal columns such as Cut or Color was performed OrdinalEncoder, on numerical columns, we standard scaled the value and the categorical columns were transformed thanks to OneHotEncoder. After encoding our dataset, we were left with high-dimensional data, which makes visualizing clusters challenging. Given that we knew some features were related, we decided to perform Principal Component Analysis (PCA). PCA reduces the dimensionality of the model while preserving maximum variance, which helps us understand the most important features and capture the underlying structure and correlations in the data. To select the best number of components for the PCA, we plotted a graph explaining variance by features, wanting to explain approximately 70% of the variance, we chose 4 components (see graph: Explained Variance by Features). Remarkably, the most important features of the principal components were related to the 4Cs: coming first Color, then Clarity, thirdly was all the numerical features constituting an ideal Cut – Depth%, Table%, Crownangle – , then last came the Price, Carat and L/W ratio.
After performing PCA, we used different clustering algorithms to divide our diamonds. For each algorithm, we first found the best parameters, then performed the algorithms on the dimensionally reduced data before plotting in 3D the clusters based on the first 3 components of the PCA. After that, we performed internal evaluation thanks to silhouette scores and sum of squared errors (SSE).
For this step, we performed KMeans algorithm thinking it could fit our data that was exhibited before as well separated in multiple columns. KMeans partitions the dataset into clusters by iteratively assigning data points to the nearest cluster centroid and updating this one to minimize within-cluster variance. To find the best number of clusters we used the Elbow Method graph. Then, we also tried Agglomerative clustering algorithm that constructs a hierarchical clustering structure by iteratively merging the closest clusters, thinking there could be some kind of hierarchy between our clusters, this was based on the assumption that diamonds with similar characteristics, such as cut, color and clarity might form natural groupings. For this one, to evaluate the best number of clusters, we used a for-loop that performed silhouette scores for a range of number of clusters. Then, we also decided to try the DBSCAN clustering algorithm being particularly effective when dealing with clusters of varying density. Indeed, diamonds could form dense clusters based on particular features like carat or price. For this algorithm we built a function to find the best parameters epsilon and minimum sample size. The epsilon parameter determines the maximum distance between two samples for them to be considered in the same neighborhood, while the minimum sample size controls the number of samples in a neighborhood for a point to be considered the core point. In the next section we will discuss the different experiments performed and the results we got.





Experiments / Results / Discussion

In this section, we delve into the experimental design to provide insights into the clustering analysis conducted on the dataset. As said before, for each clustering algorithm we performed parameter tuning to find the best ones. Thanks to the ElbowMethod, we found that 5 clusters would be optimal for the KMeans algorithms, potentially creating a balance between capturing meaningful distinctions in the data while avoiding excessive fragmentation. For Agglomerative clustering, 2 clusters were optimal, based on the for-loop used, potentially indicating that the dataset may exhibit distinct, overarching patterns. Finally, for the DBSCAN algorithm, the optimal epsilon was 1.1 with a minimum sample of 8, suggesting that the dataset contains clusters with a certain degree of density and separation. To perform parameter tuning, we decided to use the original dataset without dimensionality reduction to ensure maximum variance and maintain interpretability, indeed, interpretability of the clustering results is crucial therefore determining the number of clusters based on the original features was important to get aligned with the domain knowledge results.

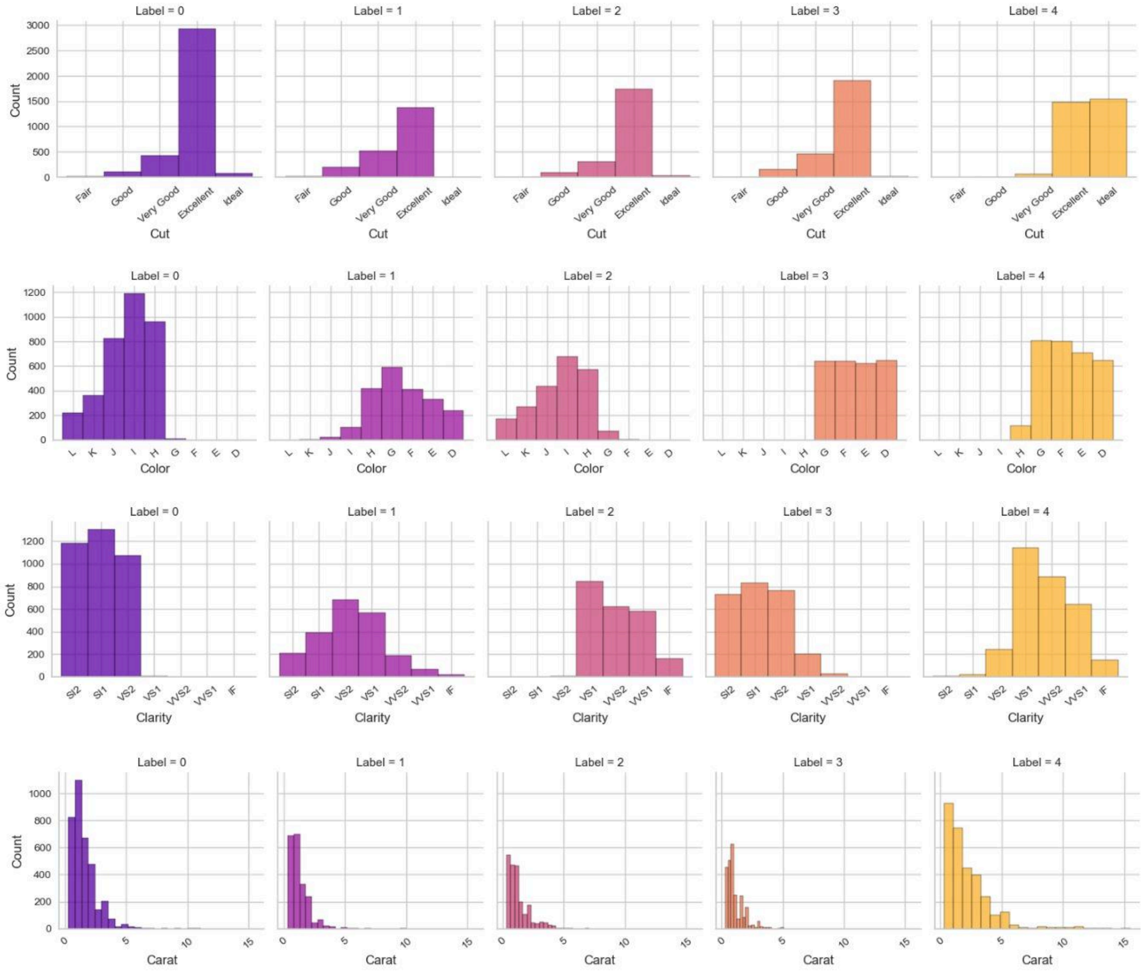
However, for the next step of performing models, we decided to fit our models on the dimensionally reduced data thanks to PCA. We chose to do so to ensure visualization, enhance performance thanks to mitigating the effect of the curse of dimensionality and also to reduce computational complexity of the models.

For the evaluation metrics, we used silhouette scores and within cluster SSE. For the KMeans algorithm, we got a silhouette score of approximately 0.234 and a SSE of 74,091. For the Agglomerative algorithm we also got a silhouette score of 0.233 but a SSE of 120,355. Finally for the DBSCAN algorithm, we got a silhouette score of 0.338 and a SSE of 238,466. Since the metrics were not sufficient to understand which model was the best, we also used a silhouette plot to understand how much diamonds were in each cluster, and we used visualization of diamonds in a 3D plot based on the first 3 PCA components. Using these methods helped us understand that for the DBSCAN algorithm, clusters were not well defined and since the SSE of Agglomerative clustering method was significantly higher than KMeans, we chose KMeans as the best model to continue to our next step of evaluating cluster characteristics and therefore incorporated a new column named Label with the cluster number for each diamond.

The KMeans algorithm effectively partitioned the diamond dataset into five distinct clusters, each exhibiting a similarity in size. This uniformity in cluster sizes (between 2100 and 3500) suggests a balanced distribution of diamonds among the identified clusters.

We plotted the different columns for each cluster to evaluate which features were the most explaining the clusters. We found that Cluster number 4 included most of the Lab diamonds, and that it usually had the best colors, clarity and cut, as well as excellent symmetry and polish. Overall the quality of diamonds in this cluster were the best, and we can explain that because half of the cluster are lab diamonds, that are created by choosing color and clarity features. In the rest of the clusters, usually they are Natural diamonds. For Cluster number 3, it also had the best colors but this time with less good clarity and usually very good or excellent cut, but not always excellent symmetry. Cluster number 2 has superior category clarity but less good colors, cut also was usually excellent. Cluster number 0 usually has a good or excellent cut, less good colors and the diamonds have bad clarity. To our understanding, this cluster includes diamonds with bad quality from the naturally found features, therefore they are cutted in excellent ways to increase their value. Finally Cluster number 1 is the cluster with the less similar characteristics for the 4Cs because of the distributions of the columns making it difficult to understand its characteristics. However we see a tendency of less good clarity but better colors with multiple sorts of cuts, polish and symmetry.

Based on the results of the clusters, we were overall satisfied with the results of the KMean algorithm, it seems to differentiate well between the different qualities of diamonds and we were able to draw effective conclusions from the results. Furthermore, we consulted experienced professionals in the industry who have over 20 years of hands-on experience. They reviewed our findings and the conclusions we drew from the results. Their deep knowledge and extensive time in the field provided valuable feedback, confirming the reliability and a certain significance from our analysis.



Conclusion and Future Work

To conclude, our study has demonstrated the effectiveness of the KMeans clustering algorithm in partitioning the diamond dataset into distinct clusters based on various attributes such as cut, color, clarity, and other characteristics. Through analysis and interpretation of the clusters, we have gained valuable insights into the qualities and distribution patterns of diamonds.

Our findings could be important for various stakeholders in the diamond industry, including manufacturers, retailers, and consumers. By understanding the clustering patterns, industry professionals can make informed decisions regarding diamond sourcing, diamond creation, pricing strategies, and market segmentation. Moreover, consumers can easily access and understand the different characteristics beyond the 4Cs to choose their diamonds and understand their value.

Looking ahead, several potential future exploration and research could be performed, such as assigning grades or scores to the diamonds taking their different characteristics into consideration to assess their quality. In addition, one potential direction could be to use computer vision techniques to evaluate inclusions and blemishes inside the diamonds. Another direction could be to perform algorithms on other shapes of diamonds other than Round Brilliant such as Pear or Emerald cuts.

For future research, investigating the temporal dynamics of diamond characteristics and market trends, or exploring the relationship between diamond clusters and consumer preferences through market research and consumer surveys could offer valuable insights for retailers and marketers.

Overall, this project serves as a foundation for future studies in the field of diamond analysis and market segmentation.

Contributions

In this section, we will outline the individual contributions of our team members.

Team member 1 – Sarah Hazziza:

- Took the lead in collecting data from online sources through web crawling.
- Played a big role in cleaning up the data and exploring it alongside Team Member 2.
- Helped build the data preprocessing pipeline and worked on feature engineering.
- Worked with Team Member 2 on using Principal Component Analysis (PCA) to reduce dimension.
- Managed algorithms' development and visualizations, creating clear visuals of the clusters and analysis.
- Analyzed KMeans clustering results and helped interpret them.
- Wrote parts of the final report and presentation, explaining tasks and analyses in detail.

Team member 2 – Niv Levi:

- Found relevant website and field of analysis, and used subject expertise to guide the project's direction.
- Led the way in cleaning and exploring the data, collaborating closely with Team Member 1.
- Built the data preprocessing pipeline, handling categorical, ordinal variables, and scaling numerical features.
- Played a big part in choosing and using PCA for dimension reduction, alongside Team Member 1.
- Managed the process of extracting results, pulling out key insights from the clustering and analysis.
- Worked with Team Member 1 to analyze clustering results and draw meaningful conclusions.
- Wrote sections of the final report, offering insights into tasks and analysis implications.

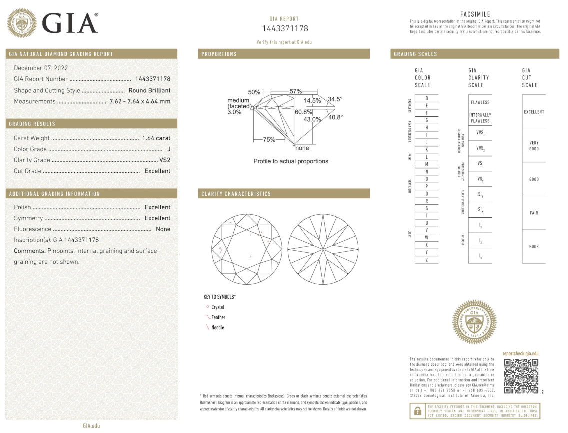
By breaking down each team member's role, we ensure transparency and acknowledge the different skills and expertise each member brought to the project. This way, we stay accountable and recognize the combined effort that made the project successful.

Appendices

Crawled website: Engagement Rings & Loose Diamonds Houston | Whiteflash

Gemological Institute Of America | All About Gemstones - GIA
International Gemological Institute | Jewelry & Gemstone Grading (igi.org)

GIA and IGI Certificate examples:



GIA

GIA NATURAL DIAMOND GRADING REPORT

December 07, 2022
GIA Report Number 1443371178
Shape and Cutting Style .... Round Brilliant
Measurements .... 7.62 - 7.64 x 4.64 mm

GRADING RESULTS

Carat Weight .... 1.64 carat
Color Grade .... J
Clarity Grade VS2
Cut Grade .... Excellent

ADDITIONAL GRADING INFORMATION

Polish .... Excellent
Symmetry .... Excellent
Fluorescence .... None
Inscription(s): GIA 1443371178
Comments: Pinpoints, internal graining and surface graining are not shown.

GIA.edu

GIA REPORT
1443371178
Verify this report at GIA.edu

PROPORTIONS

Profile to actual proportions

CLARITY CHARACTERISTICS

KEY TO SYMBOLS*
Crystal
Feather
Needle

* Red symbols denote internal characteristics (inclusions). Green or black symbols denote external characteristics (blemishes). Diagram is an approximate representation of the diamond, and symbols shown indicate type, position, and approximate size of clarity characteristics. All clarity characteristics may not be shown. Details of finish are not shown.

FACSIMILE
This is a digital representation of the original GIA Report. This representation might not be accepted in lieu of the original GIA Report in certain circumstances. The original GIA Report includes certain security features which are not reproducible on this facsimile.

GRADING SCALES

GIA COLOR SCALE
COLORLESS: D, E, F
NEAR COLORLESS: G, H, I, J
FAINT: K, L, M
VERY LIGHT: N, O, P, Q, R
LIGHT: S, T, U, V, W, X, Y, Z

GIA CLARITY SCALE
FLAWLESS
INTERNALLY FLAWLESS
VERY VERY SLIGHTLY INCLUDED: VVS$_1$, VVS$_2$
VERY SLIGHTLY INCLUDED: VS$_1$, VS$_2$
SLIGHTLY INCLUDED: SI$_1$, SI$_2$
INCLUDED: I$_1$, I$_2$, I$_3$

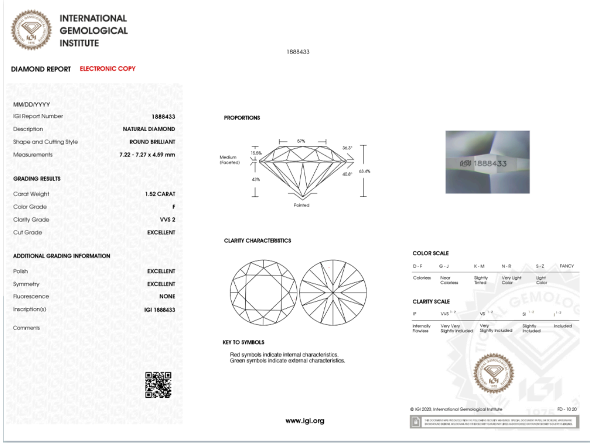
GIA CUT SCALE
EXCELLENT
VERY GOOD
GOOD
FAIR
POOR

reportcheck.gia.edu

The results documented in this report refer only to the diamond described, and were obtained using the techniques and equipment available to GIA at the time of examination. This report is not a guarantee or valuation. For additional information and important limitations and disclaimers, please see GIA.edu/terms or call +1 800 421 7250 or +1 760 603 4500. ©2022 Gemological Institute of America, Inc.

2

THE SECURITY FEATURES IN THIS DOCUMENT, INCLUDING THE HOLOGRAM, SECURITY SCREEN AND MICROPRINT LINES, IN ADDITION TO THOSE NOT LISTED, EXCEED DOCUMENT SECURITY INDUSTRY GUIDELINES.



INTERNATIONAL GEMOLOGICAL INSTITUTE

DIAMOND REPORT ELECTRONIC COPY

1888433

MM/DD/YYYY
IGI Report Number 1888433
Description NATURAL DIAMOND
Shape and Cutting Style ROUND BRILLIANT
Measurements 7.22 - 7.27 x 4.59 mm

GRADING RESULTS

Carat Weight 1.52 CARAT
Color Grade F
Clarity Grade VVS 2
Cut Grade EXCELLENT

ADDITIONAL GRADING INFORMATION

Polish EXCELLENT
Symmetry EXCELLENT
Fluorescence NONE
Inscription(s) IGI 1888433
Comments

PROPORTIONS

CLARITY CHARACTERISTICS

KEY TO SYMBOLS

Red symbols indicate internal characteristics.
Green symbols indicate external characteristics.

www.igi.org

COLOR SCALE

D - F	G - J	K - M	N - R	S - Z	FANCY
Colorless	Near Colorless	Slightly Tinted	Very Light Color	Light Color	

CLARITY SCALE

IF	VVS $^{1-2}$	VS $^{1-2}$	SI $^{1-2}$	I $^{1-2}$
Internally Flawless	Very Very Slightly Included	Very Slightly Included	Slightly Included	Included

© IGI 2020, International Gemological Institute FD - 10 20

THIS DOCUMENT WAS PRODUCED WITH THE FOLLOWING SECURITY MEASURES: SPECIAL DOCUMENT PAPER, INK SCREENS, WATERMARK BACKGROUND DESIGNS, HOLOGRAM AND OTHER SECURITY FEATURES NOT LISTED AND DO EXCEED DOCUMENT SECURITY INDUSTRY GUIDELINES.