Datasheet for 'Diamond dataset'*

Yawen Tan

December 3, 2024

Extract of the questions from Gebru et al. (2021). The dataset discussed in this datasheet is 'Diamonds' (Data and (DASL) 2024).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The primary purpose of the dataset is to assess the contributions of key factors—such as carat size, diamond color, cut, and clarity—to variations in the diamond price. Additionally, it aims to provide accurate predictions of diamond prices based on these features.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was compiled by Lou Valente, associated with JMP, a statistical software suite developed by SAS Institute.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The available information does not specify any particular funding sources or grants associated with the creation of this dataset.
4. *Any other comments?*
    - This dataset is a valuable resource for statistical analysis and educational purposes, offering insights into the relationships between various diamond characteristics and their market prices. It serves as a practical example for applying multiple regression techniques and other analytical methods.

Composition

*Code and data are available at: [https://github.com/YawennnnnnTan/Analysis-of-Diamond-Price]

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances in the dataset represent individual diamonds, each characterized by various attributes such as weight, quality, and price. All instances belong to the same category (diamonds), with no distinct types beyond this.

2. *How many instances are there in total (of each type, if appropriate)?*

    - The dataset contains 2,690 instances in total, with each instance representing a unique diamond.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset is a sample of diamonds rather than a comprehensive collection of all diamonds globally. It is not explicitly stated whether the sample is random, but given that the data was collected from the internet, it is likely to be a convenience sample. Representativeness in terms of geographic coverage or market diversity is not validated.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance consists of processed features describing a diamond. The features include: Carat weigh, Cut quality, Color grade, Clarity grade, Price (in monetary units). These features allow for statistical and predictive modeling.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - Yes, the dataset includes the "price" of each diamond as the target variable. This label allows for regression analysis to predict price based on the diamond's attributes.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - There is no explicit indication of missing data in the Diamonds dataset. Each instance includes complete information for all features (carat, cut, color, clarity, and price). However, if missing values exist, they might not have been explicitly documented.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - No, the dataset does not explicitly represent relationships between instances. Each diamond is treated as an independent observation with no inherent links to other diamonds.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - The dataset does not come with predefined splits for training, validation, or testing. Users can create their own splits based on their specific analysis or modeling needs, such as stratifying by cut or price range to ensure balanced subsets.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - The dataset does not explicitly report errors or redundancies, but potential sources of noise include subjective grading (e.g., cut, color, clarity) that could vary depending on the evaluator. Additionally, data collected from the internet might introduce inconsistencies due to differences in recording or reporting standards.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained and does not rely on external resources for its attributes. There are no references to external links, archival versions, or licensing constraints for the dataset itself. However, its description notes that the data was collected from the internet, suggesting it might originally have referenced external resources no longer linked directly.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No, the dataset does not contain any confidential data. It only includes information about diamonds, such as their physical and qualitative characteristics, which are not protected by legal privilege or confidentiality agreements.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No, the dataset does not contain any offensive, insulting, or threatening content. It focuses solely on diamond properties and prices, which are neutral and non-sensitive in nature.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset does not explicitly identify human sub-populations such as age or gender. However, it does categorize diamonds by subgroups such as cut quality, color grade, and clarity grade. These attributes can be viewed as "sub-populations" within the data, describing the distribution of diamonds across different categories.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, it is not possible to identify individuals from this dataset. The data pertains to diamonds and not to any personal or identifying information about individuals.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No, the dataset does not include sensitive information. It only contains technical and market attributes of diamonds, which are not associated with personal, biometric, or other sensitive data.

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was indirectly sourced from the internet, likely collected from commercial and market listings for diamonds. It represents observable characteristics (e.g.,

carat, cut, color, clarity) and their corresponding prices. The description does not specify validation or verification processes, so the reliability of the source data remains uncertain.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data collection was done manually or via software scripts to extract information from online resources. The exact mechanisms or validation procedures used during collection are not explicitly detailed.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset appears to be a convenience sample, extracted from available online data without explicit mention of a probabilistic or systematic sampling strategy. It is unclear if the sample was designed to be representative of the entire diamond market.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The dataset was compiled by Lou Valente, associated with JMP. The details regarding additional personnel involved or compensation methods are not provided.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The timeframe for data collection is not specified. However, given that it was sourced from the internet, it is reasonable to assume the data reflects market conditions close to the period of collection, aligning roughly with the dataset's creation. Further details on the exact dates are not available.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No explicit mention of ethical review processes, such as an institutional review board (IRB), is provided for this dataset. Since the data pertains to commodities (diamonds) rather than individuals, ethical reviews were likely deemed unnecessary.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- The data was obtained via third-party sources, specifically collected from the internet (likely from online diamond retailers or market databases).

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Not applicable, as the dataset does not involve individuals or personal data

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Not applicable, as the dataset does not involve personal data or interactions with individuals.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Not applicable, as no personal data or consent process is involved.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No, such analysis is not applicable as the dataset does not involve data subjects. It focuses solely on diamonds, which are inanimate objects with no associated ethical or personal implications.

12. *Any other comments?*

    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - The dataset appears to have undergone preprocessing to ensure that each instance includes the relevant attributes (carat, cut, color, clarity, and price). Categories such as cut, clarity, and color have been labeled into discrete values to facilitate analysis. There is no indication of missing data, suggesting that missing values were

either imputed or removed. However, detailed information on the preprocessing steps is not provided.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - There is no mention of the availability of raw data. The dataset provided represents the preprocessed and cleaned version suitable for analysis.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - The dataset does not include information about the software or tools used for preprocessing. It is likely that manual curation or commonly available statistical tools were employed, but no specific tools are documented.

4. *Any other comments?*

    - The dataset is user-ready and appears to be cleaned and well-structured for statistical modeling. While preprocessing details are not explicitly documented, the data format suggests that steps were taken to enhance usability.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The dataset has been widely used in educational contexts for teaching statistical techniques, such as multiple regression analysis and exploratory data analysis. It is commonly used in tutorials, workshops, and data science courses to demonstrate relationships between variables.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - There is no dedicated repository linking to all works or systems that have used the dataset. However, it is frequently referenced in statistical software packages like R (e.g., ggplot2) and appears in various online educational materials and textbooks

3. *What (other) tasks could the dataset be used for?*

    - Potential tasks include: Feature engineering exercises to create new attributes or improve existing models; Clustering or segmentation analyses to identify groups of similar diamonds.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues)*

or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

- The dataset's reliance on data collected from the internet introduces potential biases, as it may not represent the entire diamond market or regional variations. Users should consider the potential limitations in representativeness and avoid drawing overly general conclusions about the global diamond market. This can be mitigated by combining this dataset with more comprehensive or diverse data sources.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - The dataset should not be used for: 1. Drawing inferences about human populations, as it is unrelated to individuals or social demographics. 2. Legal or financial decisions involving real-world diamond markets without validating its representativeness or currency. 3. High-stakes decisions where data accuracy or completeness is critical, as the dataset might not capture all nuances of the diamond industry.

6. *Any other comments?*

   - No

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the dataset is publicly available and distributed via the Data and Story Library (DASL) website. It is openly accessible to anyone for educational and analytical purposes.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset is distributed through the DASL website as a downloadable file. It does not appear to have a digital object identifier (DOI).

3. *When will the dataset be distributed?*

   - The dataset is already available and has been publicly distributed for an extended period. It is accessible at any time through the DASL website.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- The dataset does not explicitly mention a copyright or specific licensing terms. However, its inclusion on the DASL website implies it is intended for free and open use, primarily for educational and research purposes. Users should verify terms of use on the DASL platform for clarity.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No known IP-based or third-party restrictions are associated with this dataset. It is freely shared for public use via DASL.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No, there are no export controls or regulatory restrictions applicable to this dataset. It is unrestricted and available globally.

7. *Any other comments?*

    - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset is hosted and maintained by the Data and Story Library (DASL), which is an educational resource for statistical data.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - There is no direct contact information provided for the dataset's owner or curator. Users can visit the DASL website for more information or to make inquiries.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - There is no specific mention of an erratum for the Diamonds dataset on the DASL website.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - There is no indication that the dataset will be updated. It appears to be static and provided as-is for educational purposes. Updates, if any, would likely need to be communicated through the DASL platform.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - Not applicable, as the dataset does not involve personal or human-related data.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Since the dataset is static and does not appear to have multiple versions, the question of supporting older versions is not applicable.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - No explicit mechanism for contributions or extensions is mentioned. Users can, however, utilize the dataset freely for their analyses and potentially create derivative datasets or studies based on it. Any such contributions would not be officially supported by DASL.

8. *Any other comments?*

    - No

References

Data, and Story Library (DASL). 2024. "Diamonds Dataset." https://dasl.datadescription.com/datafile/diamonds/?_sfm_methods=Multiple+Regression&_sfm_cases=1000+59943.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.