Kat Fritz
February 23, 2023
LIS 545 B

Data and Metadata Profile

Dataset: "Caffeine Content of Drinks—Find caffeine amounts and calories of most drink brands and types."

Dataset owner: Heiton Nunes

Stakeholders: Heiton Nunes (owner), dataset contributors, users, viewers, Caffeine Informer (source), and data analysts who have used the data. Though they may be unaware of the dataset, the businesses affiliated with the listed caffeinated beverages could be affected by the dataset.

Data Overview

"Caffeine Content of Drinks" is a dataset on the community platform Kaggle. It is owned by Heiton Nunes. The data are focused on the caffeine content in drinks. There are five attributes of the data: the drink's name, the volume quantity (ml), the calorie count, the caffeine quantity (mg), and drink type. Because most of the caffeinated beverages are manufactured, there is not set volume quantity. The source is Caffeine Informer (2023), and they state that the data were "Compiled by hand using the nutrition label, directly contacting the manufacturer, or from lab tests." There is no indication by Nunes or Caffeine Informer that any usage restrictions exist.

The data are extremely accessible. There is one comma-separated values (CSV) file for the data, making it easy for anyone to access the dataset since it does not require any specialized software to view it. Additionally, because Kaggle can display and filter the data, users can view it on the website in 3 different ways. The Compact view displays the data in a tabular view—the same view users see if they download the CSV file. The Column option provides graphs to visually represent the data and summarizing statistics. For example, it indicates there are 610 unique values, signifying there are 610 different drinks listed. One of the bar graphs shows the number of drinks

contained within caffeine quantity ranges, starting at 0-155.50 mg and increasing by 155.50 mg until 1555 mg. Lastly, The Detail option is a combination of the Compact and Column view. shows a tabular view of dataset, including all data included in the CSV file as well as statistics about the data.

Metadata

Kaggle is the host for the dataset, and it includes a section with subcategories for metadata. The metadata does not seem to adhere to a professional metadata standard. Instead, the owner included metadata structure provided by Kaggle. Though Kaggle suggests that users use the Data Package standards when uploading through their API (Public API, Kaggle), it does not look like this dataset follows Data Package standards. The metadata is not very extensive for this dataset. The authors are not provided, but it lists the owner of the dataset—Heitor Nunes—and the provenance—Caffeine Informer. Additionally, it notes that the collection methodology was through manual retrieval.

There is a summary of the dataset, indicating there is 1 CSV file comprised of 5 columns. The data are contained in 2 String columns (drink's name and drink's type), 2 Integer columns (calorie count and caffeine quantity), and 1 Decimal column (volume).

Nunes also tagged the dataset with subject: Food, Classification, Exploratory Data Analysis, and Alcohol. Including topical tags helps users locate the data. Furthermore, there is some interesting metadata about the user activity around the dataset. For example, the dataset has been viewed 17,927 times and has 2,249 downloads. This demonstrates that users have been drawn to dataset and have saved it, either to reference, look over, or even use for their own projects.

External Uses of Data

Kat Fritz
February 23, 2023
LIS 545 B

There are 7 notebooks that have been created using the data by other users in Kaggle. For example, the user Tetsuya Sasaki (2022) utilized the data to create categories within the drinks to choose which caffeinated beverage they would want to consume. The clusters they found helped them decide if a beverage had more caffeine, was more caloric, or if it had a balanced mix of caffeine and sweetness.

As for publications or works outside of Kaggle, I searched for citations that would link documents back to the dataset, but there were no returns. I used different combinations of the URL, dataset title, and contributor to search for publications using the dataset. I was unable to find literature in Google Scholar referencing the data. The only results I found through a general web search brought me to the notebooks listed on Kaggle.

Conclusion - Suggested Improvements and Enrichments

For improving data, the dataset's source has updated their list in 2023, but the dataset has not been updated in over a year. The data would be improved if the contributors or owner added more entries. Next, though there is a small section with a brief description of the data and the attributes used, the information is limited. They could add more information in the description to increase the number of times the dataset is returned when users perform searches on Kaggle or search engines.

One suggestion I have for the data itself it to add a column that contextualizes the values. The owner could include a column that calculates the percentage of the suggested daily consumption a drink has based on recommended intake. For example, the FDA (2018) recommends that on average adults should not consume more than 400mg of caffeine a day. At just 354.882 milliliters—12 fluid ounces—Black Label Brewed Coffee contains 1,555 milligrams of caffeine. That is 388.75%

of the recommended daily limit. This would add value to the dataset by simply adding a formula, allowing users to immediately see the nutritional and medical value of caffeine data. Plus, it would add onto the data retrieved from Caffeine Informer. However, this would require a different file format as CSV files cannot save formulas. Additionally, these calculations can also be done by users utilizing and referencing the dataset for new purposes.

Overall, this dataset could have many uses. It is contained in an accessible file format and is hosted on a community platform that makes the data even easier to retrieve. Adding additional columns or information could increase the probability of users interacting with and referencing the dataset, but it could also slightly diminish accessibility. It has not been updated in over a year, so the dataset could benefit from including more entries, especially if there are any new caffeinated drinks on the market.

Kat Fritz
February 23, 2023
LIS 545 B

References

Caffeine Content of Drinks. (n.d.). Caffeine Informer. Retrieved February 11, 2023, from https://www.caffeineinformer.com/the-caffeine-database

Nunes, H. (n.d.). *Caffeine Content of Drinks*. Kaggle. Retrieved February 9, 2023, from https://www.kaggle.com/datasets/heitornunes/caffeine-content-of-drinks

Public API Documentation. (n.d.). Kaggle. Retrieved February 22, 2023, from https://www.kaggle.com/docs/api

Sasaki, T. (n.d.). *Another Drink Categories by Caffeine*. Kaggle. Retrieved February 9, 2023, from https://kaggle.com/code/sasakitetsuya/another-drink-categories-by-caffeine

U.S. Food & Drug. (2021). Spilling the Beans: How Much Caffeine is Too Much? *FDA*. https://www.fda.gov/consumers/consumer-updates/spilling-beans-how-much-caffeine-too-much