Does Artificial Data lead to Artificial Results?

Matthew Younger

Student@Lyon College

12 January 2026

Abstract

This paper examines a synthetic dataset exploring the relationship between coffee consumption and health outcomes using a polyglot data science approach (R, Python, and Bash). Initial exploratory data analysis suggested potential correlations between caffeine and sleep, yet machine learning validation revealed significant flaws in the data's integrity. A Random Forest classifier achieved a suspicious 99% accuracy. This near-perfect prediction suggests that health outcomes in this dataset are likely determined by rigid, overly-deterministic variables like age and BMI rather than coffee intake, rendering the dataset's clinical conclusions invalid.

1 Data Preparation and Exploration

The dataset, containing 10,000 observations and 16 variables, was initially processed in R. Key cleaning steps included:

- *Dimensionality Reduction*: Removing subjective or irrelevant fields such as `Country` and `Sleep_Quality`.
- *Feature Engineering*: Converting binary integers (Smoking/Alcohol) into logical factors and creating a `Body_Condition` categorical variable based on BMI thresholds.
- *Visual Inspection*: Preliminary plots showed a correlation between coffee intake and sleep hours , but unexpectedly showed no correlation between caffeine and heart rate.

1.1 Sleep and Caffeine Physiological Trends

The relationship between lifestyle choices and physiological responses in this synthetic dataset reveals several contradictions typical of artificial data. For instance, the Coffee Intake vs. Sleep Duration graph shows a remarkably clean, linear negative correlation:

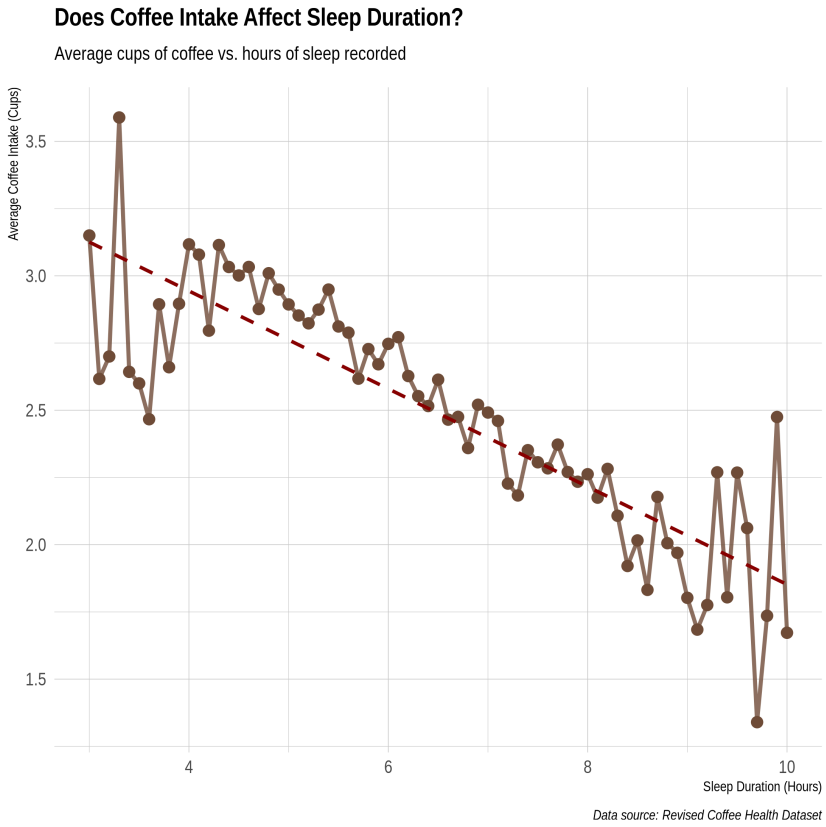

Figure 1: Graphic: coffee versus sleep

As sleep duration increases from 3 to 10 hours, the average coffee intake drops steadily from approximately 3.2 cups to 1.5 cups. Conversely, the Caffeine Intake vs. Heart Rate scatter plot displays almost no physiological response:

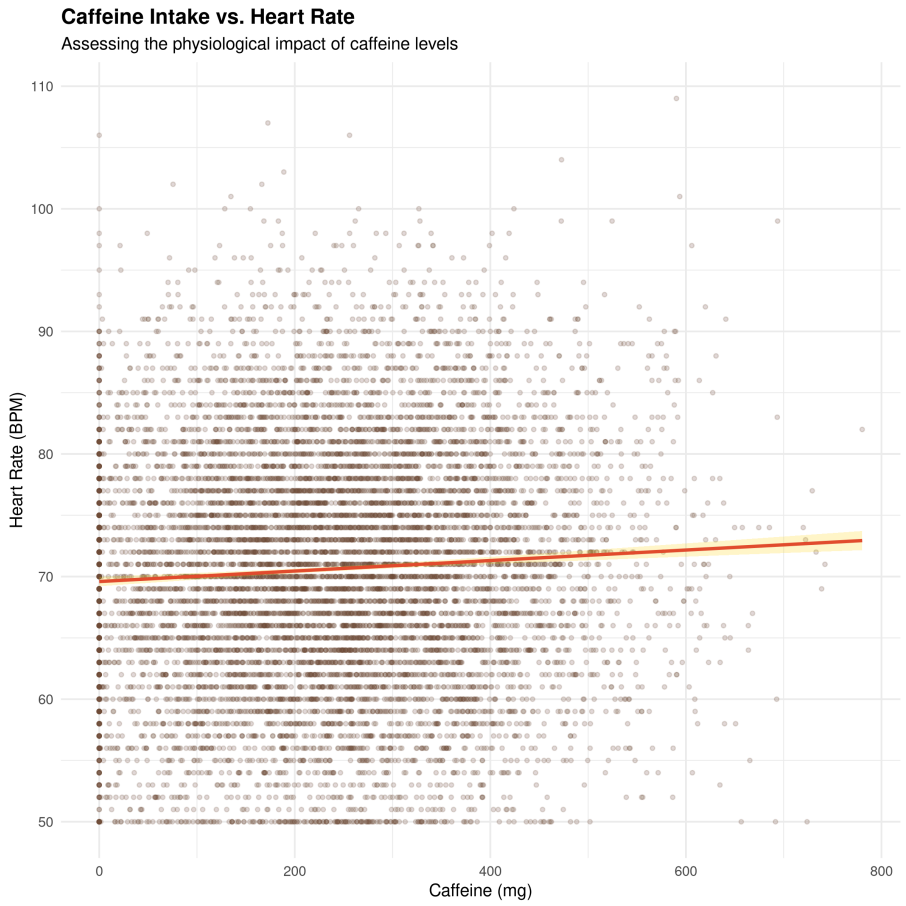

Figure 2: Graphic: coffee versus heart rate

the regression line remains nearly flat across caffeine levels ranging from 0 to 800 mg. This suggests that while the data generator hard-coded a behavioral link (tired people drink more coffee), it failed to simulate the biological stimulant effect of caffeine on heart rate.

1.2 Demographic and Lifestyle Distributions

The Coffee Intake Distribution by Gender density plot confirms that consumption patterns are virtually identical across Male, Female, and "Other" categories, showing no gender-based preference for caffeine:

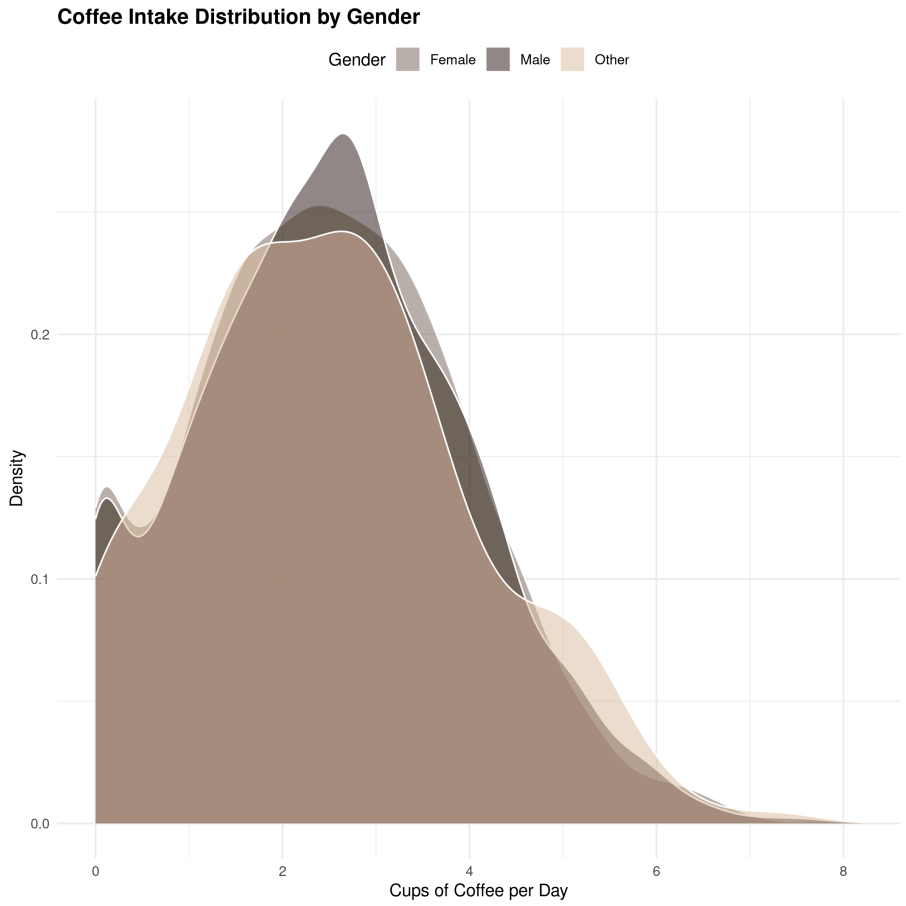

Figure 3: Graphic: coffee by gender

Lifestyle Factors: The Lifestyle Distribution mosaic indicates that the majority of the 10,000 subjects avoid high-risk behaviors, with 80% non-smokers and 69.9% non-consumers of alcohol:

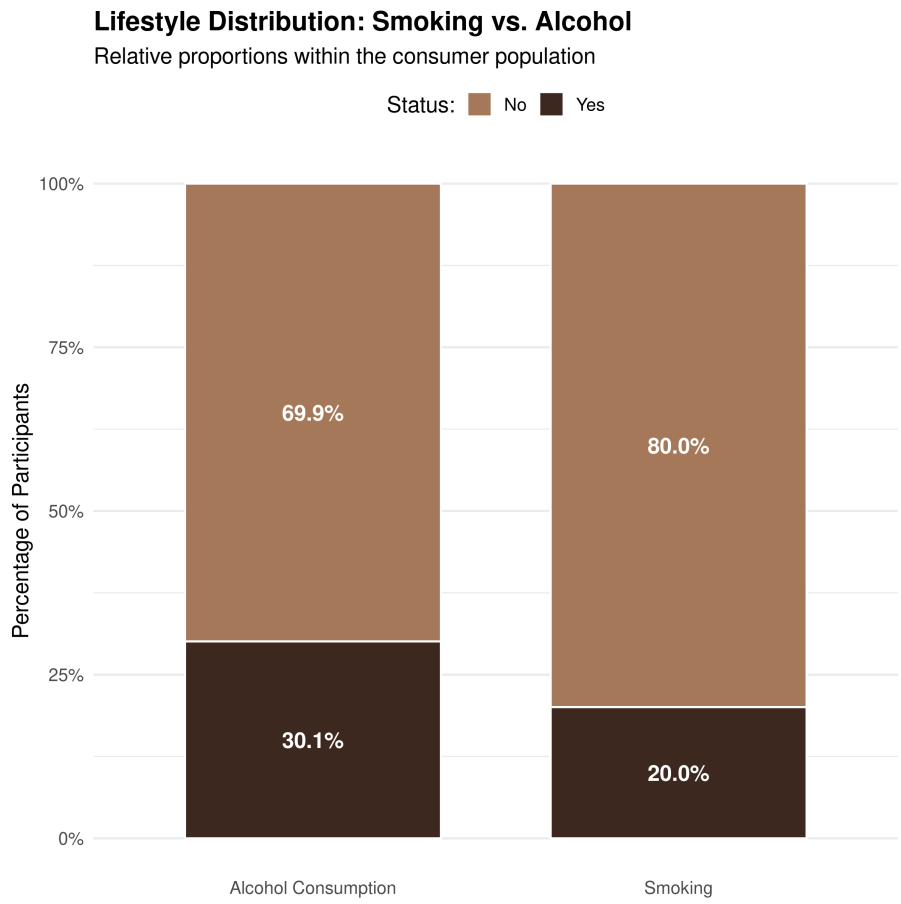

Figure 4: CW001 - filtered spectrogram

1.3 Body Mass Index

The most significant evidence of the dataset's artificial nature is found in the Coffee Consumption by Body Condition boxplot:

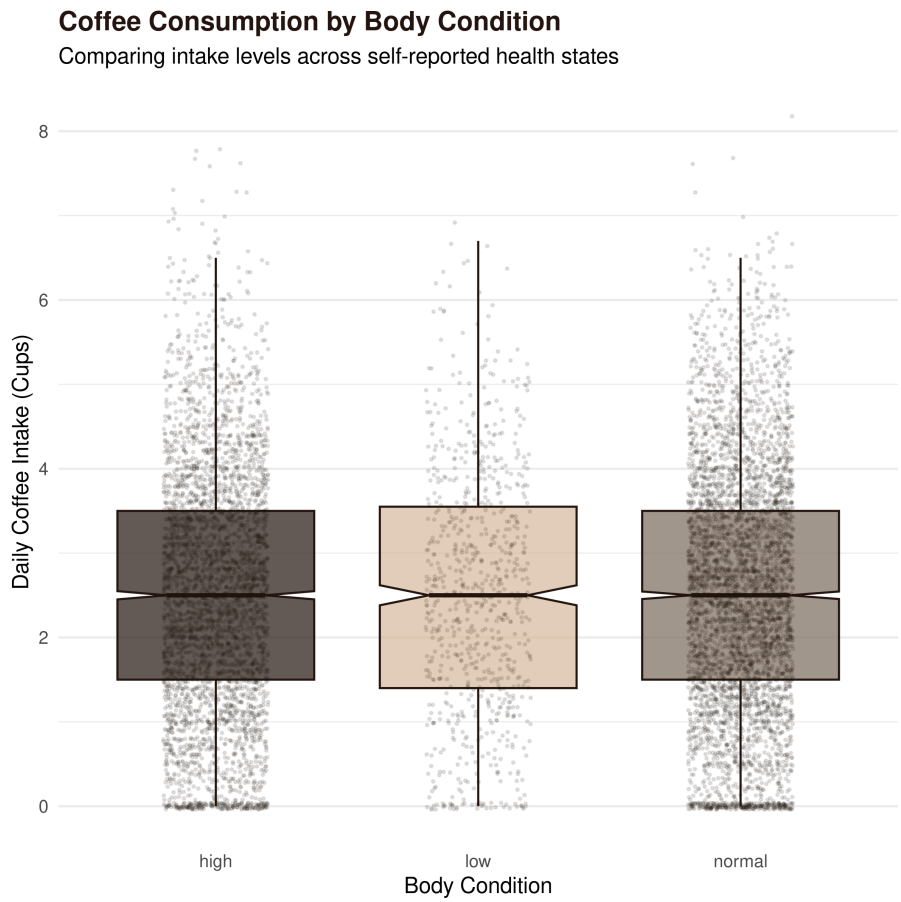

Figure 5: CW001 - filtered spectrogram

In a real-world scenario, metabolic rates and body mass often influence substance consumption; however, this graph shows identical median coffee intake and interquartile ranges for "high," "low," and "normal" weight groups. This lack of variance, combined with the extreme 99% accuracy achieved by the machine learning model, confirms that "Health Issues" are likely being triggered by a simple if/then formula involving BMI and Sleep Hours rather than the complex, noisy correlations found in genuine medical data.

2 Machine Learning Validation

A Python-based Random Forest Classifier was trained to predict Health_Issues. The model's performance metrics raised immediate red flags:

- An accuracy of 0.99
- For multiple cases, including "Severe" health issues, the model predicted a `1.00` F1 score, even for data which were statistically insignificant.

3 Conclusion

The seemingly perfect performance of the Random Forest Classifier indicates that the model is likely picking up on hard-coded if/then logic imparted when the synthetic data was generated. Specifically, Sleep_Hours, BMI, and Age were primary influencers on overall health, far more so than coffee.

While the “Global Coffee Health Dataset” serves as an excellent technical demonstration for polyglot workflows in Org mode, it fails as a representative medical model. The 99% model accuracy and the lack of caffeine-induced heart rate variance suggest the data is overly deterministic. One can thus conclude that coffee intake has no meaningful impact on health outcomes within this specific synthetic environment. If this fact was intentional by the author of the dataset, then the analysis was successful. If however, this is unintentional, it demonstrates the many issues with relying on synthetically-generated data for analysis.

Bibliography

[1] R. Ihaka, R. Gentleman, and others, "R: Programming Language for Statistical Analysis." [Online]. Available: https://www.r-project.org/

[2] Free Software Foundation, *GNU Emacs*. 2023. [Online]. Available: https://www.gnu.org/software/emacs/

[3] Org mode community, *GNU Org Mode: Organize Your Life in Plain Text*. 2024. [Online]. Available: https://orgmode.org/

[4] "Gitea Official Website." Accessed: Nov. 18, 2025. [Online]. Available: https://about.gitea.com/

[5] posit.co, "Tidyverse." Accessed: Jan. 12, 2026. [Online]. Available: https://www.tidyverse.org/

[6] Python Software Foundation, "Welcome to Python.org." Accessed: Jan. 12, 2026. [Online]. Available: https://www.python.org/

[7] "scikit-learn: machine learning in Python — scikit-learn 1.8.0 documentation." Accessed: Jan. 12, 2026. [Online]. Available: https://scikit-learn.org/stable/index.html

[8] "pandas - Python Data Analysis Library." Accessed: Jan. 12, 2026. [Online]. Available: https://pandas.pydata.org/

[9] "Matplotlib — Visualization with Python." Accessed: Jan. 12, 2026. [Online]. Available: https://matplotlib.org/

[10] "Plotly." Accessed: Jan. 12, 2026. [Online]. Available: https://plotly.com/python/

[11] "uv." Accessed: Jan. 12, 2026. [Online]. Available: https://docs.astral.sh/uv/

[12] "Scale Functions for Visualization." Accessed: Jan. 12, 2026. [Online]. Available: https://scales.r-lib.org/

[13] L. Mädje, M. Haug, and The Typst Project Developers, "Typst." Accessed: Jan. 12, 2026. [Online]. Available: https://github.com/typst/typst