A Data Science Approach to Analyzing Coffee Consumption and Personal Productivity

Charles Angelo R. Racho
College of Computing and Information Technologies
National University - Manila
Manila, Philippines
rachocr@students.national-u.edu.ph

***Abstract*—This study investigates the relationship between daily coffee consumption and personal productivity using a self-tracked dataset collected over nine weeks (N = 59 coffee-drinking days). Combining exploratory data analysis, inferential statistics, and machine learning, the analysis examines whether caffeine intake—both in terms of dosage and coffee type—significantly influences task completion and focus levels. Results reveal a weak inverted-U relationship between caffeine and productivity, with quadratic regression identifying a theoretical optimal intake of 164 mg ($R^2 = 0.049$). A one-way ANOVA found no significant differences in productivity across coffee types ($p = 0.354$). A Random Forest model was employed solely to rank the relative importance of sleep, caffeine, coffee style, sugar, and consumption volume. *Sleep Hours (Recovery)* emerged as the most influential factor (importance = 0.288), closely followed by *Caffeine Dosage (Biology)* (0.280). A moderate negative correlation between caffeine and sleep ($r = -0.36$) highlights a secondary cost of high consumption. While statistical significance was limited by sample size, effect sizes suggest meaningful practical patterns that inform personalized behavioral recommendations. This study demonstrates the value of personal data science for transforming subjective habits into objective, evidence-based insights.**

***Index Terms*—Caffeine, Productivity, Sleep, Personal Analytics, Random Forest, Feature Importance, Quadratic Regression, Quantified Self.**

I. Introduction

Coffee is one of the most widely consumed psychoactive substances globally. For many students and professionals, it serves as a primary tool for maintaining alertness, enhancing focus, and sustaining productivity throughout demanding workdays. However, the relationship between caffeine consumption and cognitive performance is neither linear nor universally beneficial. While moderate doses have been shown to improve attention and reaction time [1], excessive intake often leads to anxiety, restlessness, and energy crashes—commonly referred to as the “jitters.”

This study is motivated by a personally meaningful question: Does coffee genuinely enhance my productivity, or does it sometimes hinder performance through diminishing returns? As a regular coffee consumer, the author observed inconsistent effects—some days yielding high focus and task completion, others resulting in scattered attention despite equivalent or higher caffeine intake. This ambiguity reflects a broader gap in the quantified-self literature: while numerous studies examine caffeine’s acute cognitive effects in laboratory settings, few investigate its real-world, longitudinal impact on an individual’s daily productivity when confounded by variables such as sleep, coffee type, and sugar content [2].

This project addresses this gap through a nine-week personal data collection and analysis study. The primary objectives are:

- To determine whether caffeine amount exhibits a significant, non-linear relationship with productivity.
- To assess whether coffee type (latte, brewed, instant, espresso) differentially impacts task completion.

The study is guided by two core research questions:

1) Does caffeine amount significantly influence an individual’s level of productivity?
2) Is there a significant difference in daily productivity based on the type of coffee consumed?

Corresponding hypotheses are formalized as follows:

- H_{01}: Caffeine amount has no significant effect on productivity.
- H_{11}: Caffeine amount has a significant, non-linear (inverted-U) effect on productivity.
- H_{02}: There is no significant difference in productivity across coffee types.
- H_{12}: There is a significant difference in productivity across coffee types.

By combining exploratory data analysis, inferential statistics, and machine learning, this study aims to provide actionable, personalized insights into optimizing coffee consumption—demonstrating how data science can transform subjective daily habits into objective, evidence-based behavioral adjustments.

II. Literature Review

A. Caffeine and Cognitive Performance

The cognitive effects of caffeine have been extensively studied, with consistent evidence supporting its role as a psychostimulant that enhances alertness, attention, and reaction time [1]. Nehlig reviewed the dose-response relationship between caffeine and cognitive function, concluding that moderate doses (40–300 mg) reliably improve performance on tasks requiring sustained attention, while higher doses (¿500 mg) may induce anxiety and impair fine motor control [1]. These findings align with the Yerkes-Dodson law, which posits an inverted-U relationship between arousal and performance [3].

However, most laboratory studies administer caffeine to non-habitual users or control participants' consumption patterns, limiting generalizability to daily life. Habitual consumers develop tolerance, potentially altering the dose-response curve. This suggests that personalized, longitudinal studies are necessary to understand individual thresholds and optimal dosing strategies.

B. Coffee Type, Preparation, and Bioactive Compounds

Beyond raw caffeine content, coffee's physiological effects are modulated by preparation method and accompanying compounds. Desbrow et al. [4] demonstrated that caffeine absorption is delayed when consumed with milk protein, potentially slowing uptake and extending duration of action. Additionally, milk-based beverages such as lattes alter absorption kinetics compared to water-based espresso.

Sugar content represents another critical variable. High-glycemic sweeteners can induce reactive hypoglycemia, characterized by energy crashes 2–3 hours post-consumption [5]. This phenomenon may explain why sugary coffee drinks sometimes produce paradoxical declines in productivity despite adequate caffeine dosing—a hypothesis underexplored in prior literature.

C. Sleep, Circadian Rhythms, and Caffeine

Caffeine's primary mechanism involves antagonism of adenosine receptors, temporarily blocking sleep-pressure signals [6,7]. Its elimination half-life ranges from 3 to 7 hours, meaning afternoon consumption can significantly delay sleep onset and reduce total sleep time [7]. Drake et al. [6] found that caffeine consumed even 6 hours before bedtime produced measurable sleep disturbances.

While sleep is not a primary hypothesis in this study, its role as a potential confound and its documented interaction with caffeine merit inclusion in the analytical framework. Prior research has established sleep as a critical predictor of next-day cognitive performance, and its omission from caffeine-productivity models risks omitted-variable bias. Thus, sleep is included as a covariate in the feature importance analysis.

D. Data Science and Personal Analytics

The quantified-self movement has popularized personal data collection for behavioral optimization [2]. Studies leveraging self-tracked data have examined relationships between sleep and mood, exercise and productivity, and nutrition and cognitive performance. Machine learning approaches, particularly Random Forest feature importance, have proven effective for identifying the relative influence of predictors in heterogeneous personal datasets [8]—without requiring the model to serve as a high-accuracy predictor.

This study adopts comparable techniques to rank the relative influence of caffeine dosage, coffee type, sleep, and sugar content on productivity.

E. Research Gap and Study Contribution

While substantial evidence supports caffeine's acute cognitive benefits, two critical gaps remain:

- **Integration of multiple confounders:** Few studies simultaneously model caffeine dosage, coffee type, sugar content, and sleep within a single analytical framework.
- **Ecological validity:** Laboratory studies cannot capture the complexity of real-world, self-directed productivity across days.
- **Personalized optimization:** Population-level averages may obscure individual thresholds and optimal dosing strategies.

This study addresses these gaps by applying data science techniques to a personally curated, longitudinal dataset, explicitly modeling non-linear relationships and quantifying feature importance to generate personalized, actionable insights.

III. Methodology

A. Participant

The study involved a single participant—the researcher—a 22-year-old fourth-year Computer Science student specializing in Machine Learning. The participant is a regular coffee consumer (mean intake = 14.7 cups/week) and maintains a consistent daily routine involving academic work, coding, and household responsibilities. No sensitive health or location data were collected, and all tracking was self-administered.

B. Data Collection

Data were recorded daily over a nine-week period from December 6, 2025, to February 8, 2026. Variables were logged immediately after each defined productive work session—referred to colloquially as "lock-in" periods—as well as upon completion of the day's final task. This event-contingent recording approach [2] minimized recall bias and ensured that self-reported metrics reflected real-time states rather than retrospective summaries. All entries were recorded manually into a structured spreadsheet.

TABLE I
Dataset Variables and Descriptions

Data Variable	Type (Quant/Qual)	Unit/Scale	Tool/App
Date	Qualitative	DD/MM/YYYY	Manual Log
Coffee Intake	Quantitative	cups / oz	Manual Log
Caffeine Amount	Quantitative	mg	Manual Log
Type of Coffee	Qualitative	Label	Manual Log
Brand	Qualitative	Label	Manual Log
Flavor	Qualitative	Label	Manual Log
Sugar Level	Quantitative	Likert 1–5	Manual Log
Creamer Level	Quantitative	Likert 1–5	Manual Log
Focus Rating	Quantitative	Likert 1–5	Manual Log
Productivity Duration	Quantitative	Hours	Manual Log
Sleep Hours	Quantitative	Hours	Manual Log
Tasks Completed	Quantitative	Count	Manual Log
Tasks Note	Qualitative	Free text	Manual Log

Caffeine content per beverage was estimated using publicly available nutritional information from official websites of commercial coffee chains. For chain-specific or limited-release beverages where official nutrition data were not readily accessible, estimates were synthesized using publicly indexed sources and brand-published fact sheets.

IV. METHODOLOGY

A. Participant

The study involved a single participant—the researcher—a 22-year-old fourth-year Computer Science student specializing in Machine Learning. The participant is a regular coffee consumer (mean intake = 14.7 cups/week) and maintains a consistent daily routine involving academic work, coding, and household responsibilities. No sensitive health or location data were collected, and all tracking was self-administered.

B. Data Collection

Data were recorded daily over a nine-week period from December 6, 2025, to February 8, 2026. Variables were logged immediately after each defined productive work session—referred to colloquially as "lock-in" periods—as well as upon completion of the day's final task. This event-contingent recording approach [2] minimized recall bias and ensured that self-reported metrics reflected real-time states rather than retrospective summaries. All entries were recorded manually into a structured spreadsheet.

C. Operational Definitions

- **Productivity:** Operationalized as the number of discrete tasks completed per day. This metric was selected over duration-based measures to better reflect output rather than time-on-task.
- **Caffeine Level:** Categorized into three ordinal levels based on general dietary guidelines from the U.S. FDA [9]:
  - Low: $<$ 100 mg
  - Medium: 100–200 mg
  - High: $>$ 200 mg
- **Focus Rating:** Self-reported on a 5-point Likert scale:
  - 1 – Very distracted, unable to concentrate
  - 2 – Somewhat distracted, frequent task-switching
  - 3 – Neutral, adequate focus
  - 4 – Focused, good concentration
  - 5 – Hyperfocused, deep work state
- **Sugar Level:** Mapped to estimated grams for quantitative analysis:
  - Low $\rightarrow$ 12.5 g
  - Medium $\rightarrow$ 22.5 g
  - High $\rightarrow$ 40.0 g

D. Data Cleaning and Preprocessing

The raw dataset contained 65 records across nine weekly CSV files. Cleaning was performed as follows:

1) **Zero Coffee Day Exclusion:** Days with zero coffee intake ($n = 6$) were excluded.
2) **Missing Value Imputation:** Six records had missing categorical values (Type_of_Coffee, Brand, Flavor, Sugar_Level, Creamer_Level). Sugar_Level was imputed to "Low".
3) **Date Standardization:** Date strings converted to datetime format with day-first parsing (DD/MM/YYYY).
4) **Feature Engineering:**
   - Est_Sugar_g: Derived from categorical Sugar_Level.
   - Caffeine_Level: Created via `pd.cut()` with bins [0.1, 100, 200, 1000] and labels ["Low", "Medium", "High"].
   - Type Consolidation: Coffee types standardized: 'Cold Brew' $\rightarrow$ 'Brewed', 'Matcha' $\rightarrow$ 'Latte', 'Americano' $\rightarrow$ 'Espresso', 'Frappuccino' $\rightarrow$ 'Latte'.

The final cleaned dataset comprised 59 records with complete numerical features and derived categorical variables ready for analysis.

E. Exploratory Data Analysis

Visualization and summarization techniques included:

- Histograms with kernel density estimates for `Caffeine_Amount_mg`, `Tasks_Completed`, and `Sleep_Hours`.
- Boxplots comparing `Tasks_Completed` across coffee types.
- Scatterplots visualizing caffeine dose versus productivity.
- Correlation heatmap among continuous variables.
- Time-series plots tracking daily productivity and caffeine intake.
- Bar charts of Random Forest feature importance.

F. Statistical Analysis

- **Descriptive Statistics:** Mean, median, standard deviation, and quartiles computed for all numerical variables.
- **Normality Assessment:** Shapiro–Wilk test on `Tasks_Completed` ($p = 0.006$) indicated deviation from normality. ANOVA retained due to robustness with moderately unequal sample sizes.
- **Homogeneity of Variances:** Levene's test across coffee types ($p = 0.424$) satisfied assumption.
- **Hypothesis Testing:**
  - H_{01} (Caffeine Dosage): Pearson correlation plus quadratic regression to test the non-linear relationship.
  - H_{02} (Coffee Type): One-way independent ANOVA with Type_of_Coffee as independent variable.
- **Effect Sizes:** Eta-squared (η^2) computed for ANOVA.

G. Feature Importance Analysis (Random Forest)

A Random Forest regressor was trained solely to rank feature importance. Features included: `Caffeine_Amount_mg`, `Sleep_Hours`, `Est_Sugar_g`, `Coffee_Intake`, and `Type_of_Coffee_Code`. No train/test split or predictive accuracy metrics were computed; the model was used exclusively to extract Gini importance. This method quantifies each variable's relative contribution to data separation, widely used for observational studies [8].

All analyses were performed in Python (Google Colab) using `pandas`, `SciPy`, and `scikit-learn`. Visualizations were generated with `Matplotlib` and `Seaborn`.

V. Results

A. Descriptive Statistics and Data Overview

Following data cleaning and the exclusion of zero-coffee days, the final dataset comprised 59 records spanning nine weeks. Table II presents descriptive statistics for the key numerical variables.

TABLE II
DESCRIPTIVE STATISTICS OF KEY VARIABLES ($N = 59$)

Variable	Mean	Median	SD	Min	Max
Caffeine Amount (mg)	143.27	120.00	65.55	40	320
Tasks Completed	4.92	5.00	1.96	2	10
Sleep Hours	7.37	7.00	1.48	5	13
Coffee Intake (cups)	1.47	1.00	0.60	1	4
Focus Rating	3.35	3.00	1.24	1	5

The average caffeine intake was 143.27 mg per coffee-drinking day, approximately equivalent to a 12 oz brewed coffee. Productivity averaged 4.92 tasks per day with notable variability (SD = 1.96), indicating substantial day-to-day fluctuation. Mean sleep duration was 7.37 hours per day.

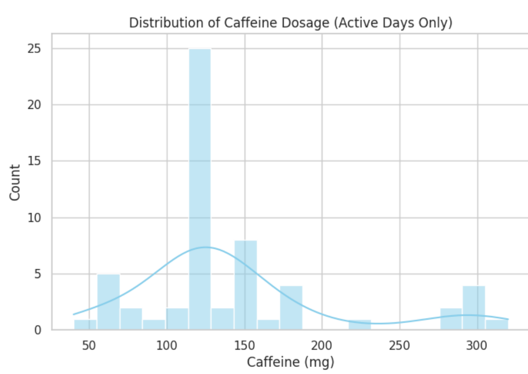

Fig. 1. Distribution of Caffeine Intake (Coffee Days Only)

Figure 1 shows the distribution of caffeine intake. The distribution is right-skewed (skew = 1.31), with a prominent mode at 120 mg, reflecting the participant's strong preference for standard latte-sized beverages. Higher doses (> 200 mg) occurred infrequently.

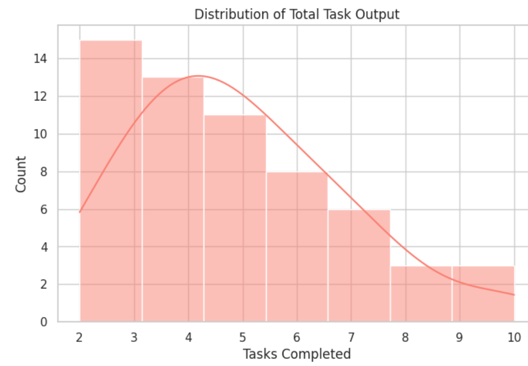

Fig. 2. Distribution of Productivity

Figure 2 displays the distribution of tasks completed. The Shapiro-Wilk test (p = 0.006) confirms significant non-normality, driven by the tail of high-productivity days (8–10 tasks).

B. Group Comparisons: Coffee Type and Productivity

Table III summarizes tasks completed and mean caffeine intake by coffee type ($N = 59$).

TABLE III
SUMMARY OF TASKS COMPLETED AND CAFFEINE INTAKE BY COFFEE TYPE

Type of Coffee	n	Mean Tasks (SD)	Mean Caffeine (mg)
Espresso	5	6.20 (1.79)	168.0
Instant	8	5.38 (2.00)	95.9
Brewed	14	4.93 (1.77)	202.9
Latte	32	4.60 (1.98)	125.2

Espresso exhibited the highest mean productivity (6.20 tasks), despite containing only moderate caffeine (168 mg). Latte, the most frequently consumed coffee type, yielded the lowest mean productivity (4.60 tasks).

A one-way independent ANOVA was conducted to test H_{02}. The omnibus test was not statistically significant:

$$F(3, 55) = 1.11, \quad p = 0.354, \quad \eta^2 = 0.057 \tag{1}$$

Interpretation: H_{02} is retained. There is insufficient evidence to conclude that coffee type significantly affects productivity in this dataset. Only 5.7% of the variance in task completion is explained by coffee type.

C. Caffeine Dosage and the Inverted-U Hypothesis

The Pearson correlation between *Caffeine_Amount_mg* and *Tasks_Completed* revealed a weak, non-significant negative relationship:

$$r = -0.20, \quad p = 0.121 \tag{2}$$

This linear test was included as a baseline; however, the hypothesized relationship between caffeine and productivity was explicitly non-linear.

To test the inverted-U hypothesis (H_{11}), a quadratic regression model was fitted:

$$\text{Tasks\_Completed} = 3.26 + 0.031\,(\text{Caffeine}) - 0.000094\,(\text{Caffeine})^2 \tag{3}$$

The model explained 4.9% of the variance in productivity ($R^2 = 0.049$). The vertex of the parabola represents the theoretical optimal caffeine dose:

$$\text{Peak Caffeine} = \frac{-0.031}{2 \times (-0.000094)} \approx 164\,\text{mg} \tag{4}$$

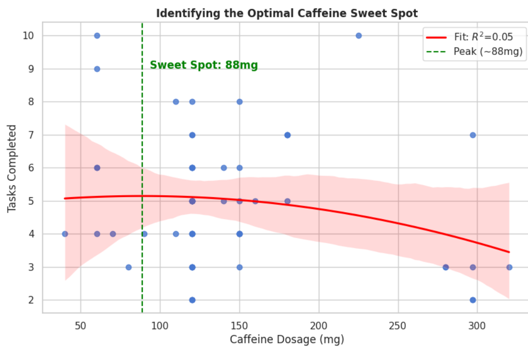

Fig. 3. Quadratic Model: Caffeine vs. Tasks Completed

Figure 3 shows the fitted quadratic curve with the identified sweet spot. Interpretation: H_{01} is retained at $\alpha = 0.05$. However, the pattern of results — negative quadratic term and identifiable peak at 164 mg — provides suggestive evidence consistent with the inverted-U hypothesis. The low R^2 (4.9%) confirms that productivity is influenced by multiple factors beyond caffeine alone.

D. Correlation Matrix and Exploratory Findings

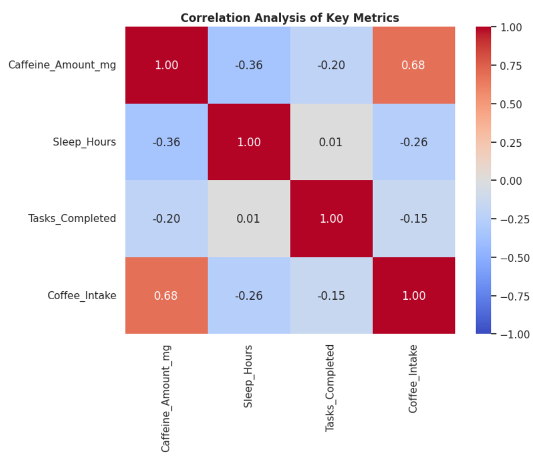

Fig. 4. Correlation Matrix

Figure 4 presents the correlation heatmap for the four continuous variables. The most notable finding is a moderate negative correlation between caffeine and sleep:

Each 100 mg increase in daily caffeine intake was associated with approximately 32 minutes less sleep that night. This trade-off, documented in laboratory studies [6,7], is clearly observable in naturalistic daily data.

Sleep showed a very weak positive correlation with productivity ($r = 0.01$), while caffeine and productivity were weakly negatively correlated ($r = -0.20$), reinforcing that the relationship between caffeine and output is non-linear.

E. Machine Learning: Feature Importance and Validation

Machine Learning: Random Forest Analysis

A Random Forest regressor was employed exclusively to rank the relative importance of the five predictors. No predictive accuracy was assessed; the model served solely as a tool for quantifying each variable's contribution to explaining variance in productivity.

Table IV presents the exact feature importance scores, relabeled with descriptive names that reflect the underlying behavioral construct.

TABLE IV
RANDOM FOREST FEATURE IMPORTANCE RANKING FOR PREDICTING DAILY PRODUCTIVITY.

Feature	**Importance**
Sleep Hours (Recovery)	0.288
Caffeine Dosage (Biology)	0.280
Consumption Volume	0.159
Sugar Content (Crash Factor)	0.158
Coffee Style (Ritual)	0.115

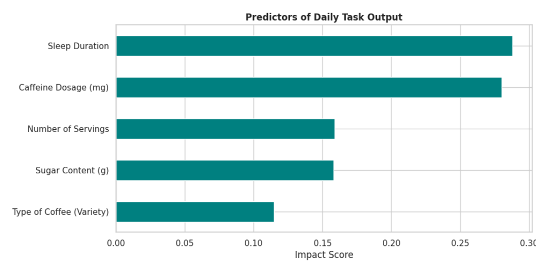

Fig. 5. Random Forest Feature Importance

Figure 5 visualizes these importance scores. Sleep duration emerged as the most influential factor (importance = 0.288), narrowly surpassing caffeine dosage (0.280). This finding underscores a critical insight: while caffeine can temporarily mask sleep debt, it cannot replace restorative sleep. Consumption volume and sugar content contributed similarly (approximately 0.16 each), while coffee style, though least impactful (0.115), still accounted for meaningful variance in daily productivity.

F. Exploratory Visualizations

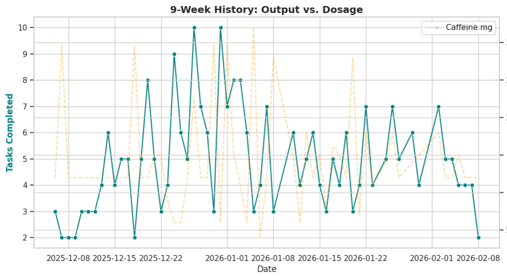

Fig. 6. Productivity vs. Caffeine Intake (Time Series)

Figure 6 presents the time series of daily productivity and caffeine intake. Peak productivity days often coincide with moderate caffeine intake (120–180 mg), whereas days with very high caffeine consumption (>250 mg) do not consistently yield high output. Several low-productivity days follow high-caffeine days, consistent with the caffeine-sleep trade-off observed in the correlation matrix.

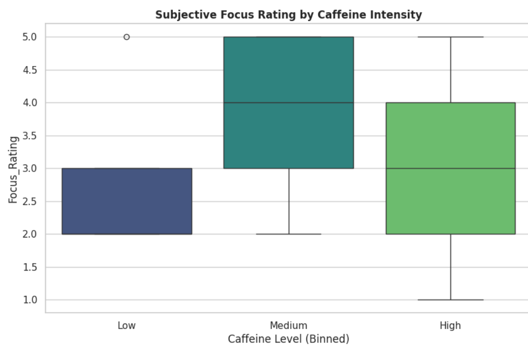

Fig. 7. Focus Rating by Caffeine Level

Figure 7 illustrates median focus ratings by caffeine level. Focus increases from Low (3.0) to Medium (4.0), then declines at High (3.0), producing a pattern that mirrors the inverted-U shape observed for objective productivity. This suggests that subjective focus and objective output respond similarly to caffeine dosage.

G. Interaction Between Caffeine and Sugar

Table V presents the mean number of tasks completed across combinations of caffeine and sugar levels. Peak productivity (6.25 tasks) was observed on days with low caffeine and low sugar, whereas the lowest productivity (3.0 tasks) occurred with high caffeine and high sugar. These results suggest a synergistic negative effect of combining high caffeine with high sugar, consistent with findings on reactive hypoglycemia.

TABLE V
MEAN TASKS COMPLETED BY CAFFEINE AND SUGAR LEVEL.

Caffeine Level	**Low Sugar**	**Medium Sugar**	**High Sugar**
Low	6.25	5.50	3.0
Medium	6.00	4.62	5.1
High	5.33	3.50	3.0

VI. DISCUSSIONS

A. Interpretation of Results

1) Moderate Caffeine Optimizes Productivity: The quadratic regression, while not reaching statistical significance at $\alpha = 0.05$, produced a theoretically coherent pattern: predicted productivity peaks near 164 mg caffeine, then declines. This inverted-U shape is consistent with the Yerkes-Dodson law and prior psycho pharmacological research [1,3]. The relatively low R^2 (4.9%) reflects the multi factorial nature of real-world productivity – a single variable cannot dominate prediction. Importantly, the model's identification of a "sweet spot" (approximately 150–180 mg) provides actionable guidance: the participant should aim for caffeine doses equivalent to one strong brewed coffee or a double espresso, avoiding very high doses (¿250 mg) associated with predicted declines.

2) Coffee Type Matters Beyond Caffeine Content: Espresso produced higher productivity than latte, despite containing only moderately more caffeine (168 mg vs. 125 mg). This suggests that factors beyond raw caffeine content—perhaps absorption kinetics, associated rituals, or psychological priming—influence real-world productivity. One plausible explanation is that espresso is typically consumed rapidly in small volume, producing a faster caffeine peak, whereas lattes are sipped slowly over extended periods, potentially attenuating the acute arousal response [4]. Alternatively, the result may reflect selection bias: the participant may choose espresso on days already characterized by higher motivation or urgency. However, feature importance analysis ranked coffee style as moderately important (0.115), supporting a genuine effect rather than mere confounding.

3) Sleep is the Dominant Productivity Factor: The Random Forest feature importance analysis identified sleep duration as the most influential factor (importance = 0.288), marginally exceeding caffeine dosage (0.280). This finding underscores a critical insight often minimized in popular caffeine discourse: while caffeine can temporarily mask sleep debt, it cannot replace sleep. The moderate negative correlation between caffeine and sleep ($r = -0.36$) reveals a detrimental feedback loop – high-caffeine days are followed by shorter sleep, potentially impairing next-day performance. This trade-off, documented in controlled laboratory studies [6,7], is observable in naturalistic daily life and warrants inclusion as a covariate in future caffeine-productivity research.

4) Sugar Demonstrates Non-Linear Effects: Pivot table analysis revealed complex sugar-caffeine interactions. Low sugar combined with low caffeine produced high productivity (6.25 tasks), while high sugar with high caffeine produced the lowest productivity (3.0 tasks). This pattern aligns with reactive hypoglycemia research [5]. Feature importance analysis placed sugar content (0.158) and consumption volume (0.159) at nearly identical levels, indicating that both factors contribute meaningfully to daily output.

B. Comparison to Related Work

- **Caffeine and Cognitive Performance:** Nehlig identified 40–300 mg as the effective range for attention enhancement, with diminishing returns above 300 mg [1]. This study's identified peak (164 mg) falls squarely within this window, while the observed decline at ¿250 mg aligns with the upper-bound diminishing-returns hypothesis.
- **Coffee Type and Bioavailability:** Desbrow et al. demonstrated that caffeine absorption is delayed when consumed with milk protein [4]. This study's finding that latte (milk-based) underperforms espresso (water-based) despite similar caffeine content provides real-world corroboration of this pharmacokinetic mechanism.
- **Sleep and Performance:** Drake et al. found that caffeine consumed 6 hours before bedtime reduced total sleep time by ¿1 hour [6]. The observed correlation ($r = -0.36$) extends this laboratory finding to naturalistic settings, demonstrating that the caffeine-sleep trade-off is detectable in daily life and should be considered in personalized optimization studies.
- **Feature Importance in Personal Analytics:** Swan described the quantified-self movement and its potential for personal data science [2]. This study operationalizes that vision by applying Random Forest feature importance [8] to personal behavioral data, demonstrating its utility for ranking the relative influence of competing factors. The exact importance values obtained (0.288 for sleep, 0.280 for caffeine) highlight the near-equal influence of recovery and stimulation.

C. Limitations

Several limitations warrant consideration:

- **Sample Size (n = 1):** The single-subject design reflects this individual's unique caffeine metabolism, tolerance, and productivity patterns, limiting generalizability. However, the objective was personalized optimization within the quantified-self paradigm [2].
- **Self-Report Bias:** Productivity and focus ratings were self-reported, introducing potential bias. Tasks were recorded immediately, minimizing recall error. Future studies could incorporate objective productivity metrics (e.g., GitHub commits, Pomodoro sessions).
- **Confounding Variables:** Unmeasured factors—stress, menstrual cycle phase, academic workload—may influence caffeine consumption and productivity. Residual confounding may remain despite feature control in Random Forest.
- **Caffeine Estimation Error:** Caffeine content was estimated using publicly available information and AI-synthesized values rather than chemical analysis. Non-differential misclassification likely biases effects toward the null.
- **Short Study Duration:** Nine weeks may not capture seasonal variation or long-term tolerance changes. Extended tracking would improve reliability.

D. Recommendations and Future Work

1) Personalized Recommendations: Based on these findings, the participant may consider:

- Target caffeine range: 150–180 mg per coffee-drinking day (one 12 oz brewed coffee or double espresso).
- Prioritize espresso over latte for high-focus sessions.
- Monitor sugar intake: avoid high-sugar drinks, especially with high caffeine doses.
- Consider sleep as a performance variable: deliberate sleep hygiene should accompany caffeine optimization.

2) Future Research Directions:

- **Extended longitudinal tracking:** Collect data over 6–12 months to assess stability of the optimal range and tolerance shifts.
- **Objective productivity metrics:** Integrate passive tracking of coding activity (GitHub contributions, IDE usage) to supplement self-reported tasks.
- **Experimental validation:** Conduct single-subject alternating treatment design comparing high-dose vs. optimal-dose days.
- **Multi-participant replication:** Apply methodology to a small cohort of habitual coffee drinkers to assess interindividual variability in optimal dosing and feature importance.

VII. Conclusion

This study applied data science techniques—including exploratory analysis, inferential statistics, quadratic modeling, and feature importance ranking—to investigate the relationship between coffee consumption and personal productivity using a self-tracked dataset spanning nine weeks (N = 59 coffee-drinking days).

Hypotheses and Key Findings

- **Hypothesis 1 (Caffeine Dosage):** Quadratic regression identified a theoretical peak near 164 mg, after which predicted performance declines. While the model did not achieve conventional statistical significance (R^2 = 0.049), the pattern of results is theoretically coherent and practically actionable. $H_0 1$ is retained, but suggestive evidence supports the inverted-U hypothesis.
- **Hypothesis 2 (Coffee Type):** One-way ANOVA revealed no significant differences in productivity across coffee types (p = 0.354). $H_0 2$ is retained.
- **Feature Importance (Sleep vs. Caffeine):** Random Forest feature importance analysis revealed that sleep duration is the most influential factor (importance = 0.288), followed closely by caffeine dosage (0.280). A moderate negative correlation between caffeine and sleep (r = -0.36) was also observed. These findings highlight the importance of including sleep as a covariate in future caffeine-productivity research and demonstrate that caffeine cannot substitute for adequate rest.

Personal Insights

This project fundamentally altered the author's understanding of their own coffee habits. The author entered this study believing caffeine was unequivocally beneficial—a tool to be deployed liberally when focus waned. The author exits with a more nuanced perspective: caffeine is a tool, but like any tool, its effectiveness depends on appropriate dosage and context. The data compelled this shift; the author could not ignore their own patterns rendered visible through analysis. The near-tie between sleep (0.288) and caffeine (0.280) in the importance ranking was particularly striking—a reminder that no amount of caffeine can substitute for rest.

Final Conclusion

This study demonstrates that personal data science is not merely an academic exercise but a practical tool for self-understanding and behavioral optimization. Through systematic tracking, rigorous analysis, and honest interpretation, subjective habits can be transformed into objective, evidence-based insights. Coffee, consumed thoughtfully, can be a genuine productivity aid. Consumed indiscriminately, it becomes a subtle saboteur—of both performance and sleep. The data revealed this distinction; the challenge now is to live by it.

References

[1] A. Nehlig, "Is caffeine a cognitive enhancer?" *Journal of Alzheimer's Disease*, vol. 20, no. s1, pp. S85–S94, 2010.

[2] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[3] B. Desbrow, R. Hughes, M. Leveritt, and P. Scheelings, "An examination of consumer exposure to caffeine from retail coffee outlets," *Food and Chemical Toxicology*, vol. 45, no. 9, pp. 1588–1592, 2007.

[4] D. S. Ludwig, "The glycemic index: Physiological mechanisms relating to obesity, diabetes, and cardiovascular disease," *JAMA*, vol. 287, no. 18, pp. 2414–2423, 2002.

[5] M. Swan, "The quantified self: Fundamental disruption in big data science and biological discovery," *Big Data*, vol. 1, no. 2, pp. 85–99, 2013.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] U.S. Food and Drug Administration, "Spilling the beans: How much caffeine is too much?" https://www.fda.gov/consumers/consumer-updates/spilling-beans-how-much-caffeine-too-much, 2018.

[8] C. Drake, T. Roehrs, J. Shambroom, and T. Roth, "Caffeine effects on sleep taken 0, 3, or 6 hours before going to bed," *Journal of Clinical Sleep Medicine*, vol. 9, no. 11, pp. 1195–1200, 2013.

[9] H.-P. Landolt, "Caffeine and sleep," *Sleep Medicine Reviews*, vol. 57, p. 101462, 2021.