Final Project Report: Analysis of Coffee Consumption and Health Metrics

Inferential Statistics 2025/2026

Andrea Recchiuti and Rafaelle Richel Pearl

2025-12-20

Contents

1 Abstract

This report investigates the relationship between caffeine consumption, lifestyle habits, and cardiovascular health using a synthetic dataset of 10,000 records. We specifically test the physiological validity of the 400mg "safe" caffeine limit using a Welch Two-Sample t-test. Furthermore, we employ Multiple Linear Regression to determine the relative impact of caffeine, sleep, BMI, and smoking on resting heart rate. Our results indicate that exceeding 400mg of caffeine significantly increases heart rate, and that sleep hours act as a significant protective factor, whereas BMI and smoking did not show significant linear effects in this specific model.

2 Introduction

1. coffee is widely used
2. 400mg rule suggested by EFSA and FDA making 400mg th daily limit for adults
3. see if this arbritrary threshold is proven by actual physiological datamention papers stating caffeine acting as stimulant on the nervous system which increases heart rate, and sleep deprivation leads to nervous system imbalance. this is done to show the possibility that caffeine is not very healthy when consumed too much.

3 Data

The primary data source for this study is the "Global Coffee Health Dataset," a synthetic population of 10,000 records designed to simulate real-world correlations between lifestyle choices and health outcomes. The dataset comprises 16 variables, detailed as follows:

1. ID (Integer): Unique identifier for each participant.
2. Age (Integer): Age group of the participant (18-80 Years).
3. Gender (Categorical): Gender (Male, Female, Other).
4. Country (Categorical): Country of residence (20 Countries)
5. Coffee_Intake (Float): Number of cups of coffee consumed daily.
6. Caffeine_mg (Float): Estimated daily caffeine intake in milligrams (1 cup is 95mg).
7. Sleep_Hours (Float): Average hours of sleep per night.
8. Sleep_Quality (Categorical): Self-reported sleep quality (Poor, Fair, Good, Excellent).
9. BMI (Float): Body Mass Index of the participant.
10. Heart_Rate (Integer): Resting heart rate in beats per minute.
11. Stress_Level (Categorical): Self-reported stress level (Low, Medium, High).
12. Physical_Activity_Hours (Float): Average hours of physical activity per week.
13. Health_Issues (Categorical): Self-reported health issues related to caffeine consumption (None, Mild, Moderate, Severe).
14. Occupation (Categorical): Type of occupation (Office, Healthcare, Student, Service, Other).
15. Smoking (Boolean): Whether the participant is a smoker (0 = No, 1 = Yes).
16. Alcohol_Consumption (Boolean): Whether the participant consumes alcohol (0 = No, 1 = Yes).

3.1 Summary of the dataset

Here is the summary of the dataset provided by the "summary()" function in R:

```
##        ID             Age           Gender            Country
##  Min.   :    1   Min.   :18.00   Length:10000       Length:10000
##  1st Qu.: 2501   1st Qu.:26.00   Class :character   Class :character
##  Median : 5000   Median :34.00   Mode  :character   Mode  :character
##  Mean   : 5000   Mean   :34.95
##  3rd Qu.: 7500   3rd Qu.:43.00
##  Max.   :10000   Max.   :80.00
##  Coffee_Intake    Caffeine_mg     Sleep_Hours     Sleep_Quality
##  Min.   :0.000   Min.   :  0.0   Min.   : 3.000   Length:10000
##  1st Qu.:1.500   1st Qu.:138.8   1st Qu.: 5.800   Class :character
##  Median :2.500   Median :235.4   Median : 6.600   Mode  :character
##  Mean   :2.509   Mean   :238.4   Mean   : 6.636
##  3rd Qu.:3.500   3rd Qu.:332.0   3rd Qu.: 7.500
##  Max.   :8.200   Max.   :780.3   Max.   :10.000
```

```
##       BMI          Heart_Rate     Stress_Level       Physical_Activity_Hours
##  Min.   :15.00   Min.   : 50.00   Length:10000       Min.   : 0.000
##  1st Qu.:21.30   1st Qu.: 64.00   Class :character   1st Qu.: 3.700
##  Median :24.00   Median : 71.00   Mode  :character   Median : 7.500
##  Mean   :23.99   Mean   : 70.62                      Mean   : 7.487
##  3rd Qu.:26.60   3rd Qu.: 77.00                      3rd Qu.:11.200
##  Max.   :38.20   Max.   :109.00                      Max.   :15.000
##  Health_Issues       Occupation        Smoking    Alcohol_Consumption
##  Length:10000       Length:10000       No :7996   No :6993
##  Class :character   Class :character   Yes:2004   Yes:3007
##  Mode  :character   Mode  :character
##
##
##
```

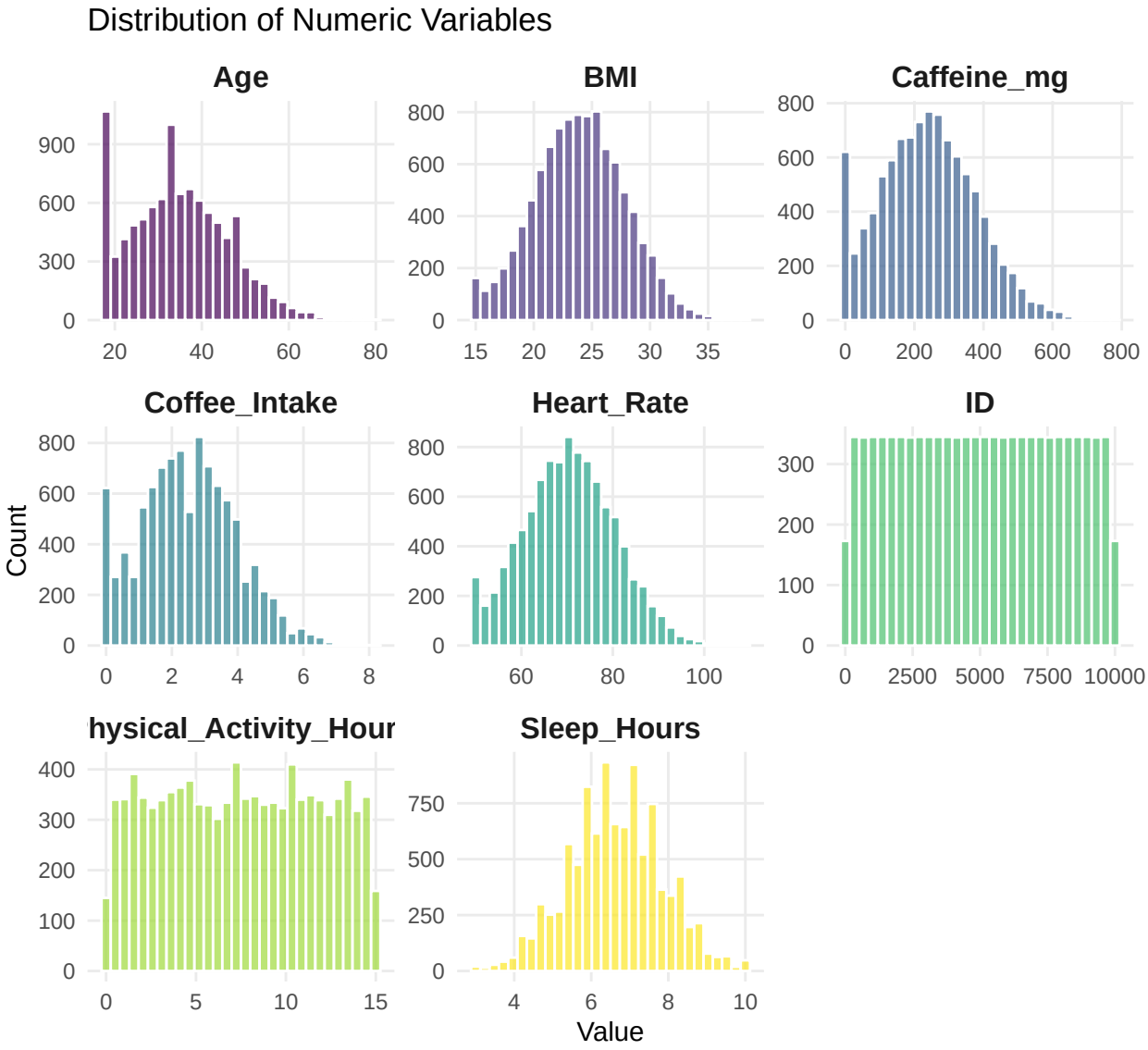
3.2 Exploratory Data Analysis

To assess the integrity of the data and verify the assumptions required for inferential testing, we conducted a systematic Exploratory Data Analysis (EDA).

3.2.1 Distribution of Numerical Variables

Continuous variables were isolated from categorical descriptors to mitigate noise during visualization. Notably, binary variables (Smoking and Alcohol_Consumption) were refactored into factors to ensure correct statistical treatment. Data was transformed from a wide to a long format to generate a high-density faceted grid of histograms.

We utilized "free scales" for the axes to prevent variables with smaller numerical ranges (e.g., BMI) from being visually suppressed by those with larger ranges (e.g., Caffeine Intake). This visual profiling allowed us to confirm the symmetry and skewness of the predictors, identify potential outliers, and ensure all ranges were physiologically plausible.



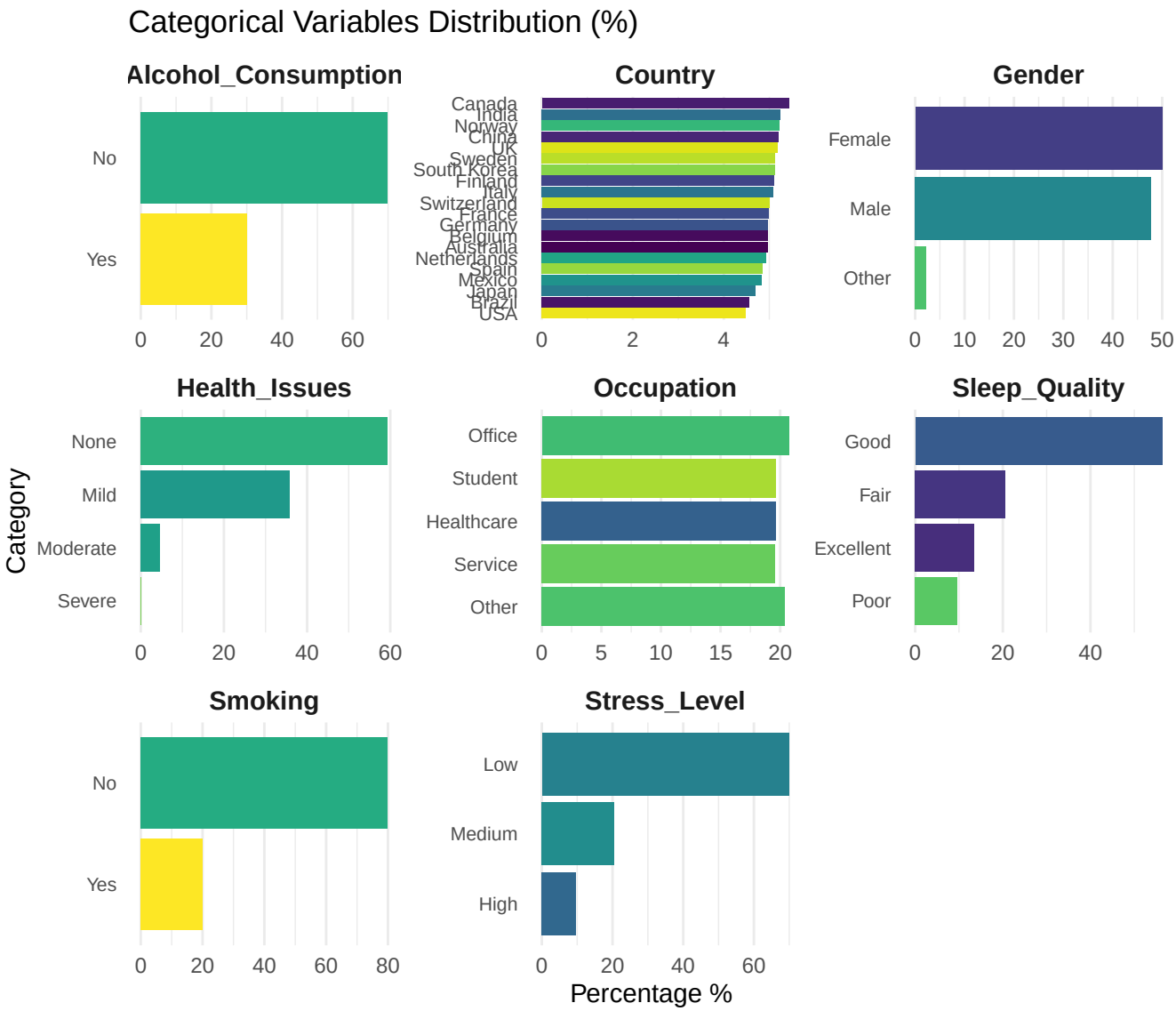
Normality & Parametric Validity: Variables like Heart_Rate and Sleep_Hours follow a clear bell-shaped curve. This is a vital finding because it satisfies the "Normality Assumption" required for the T-test and Linear Regression you performed later.

Caffeine Habits: The distribution of Caffeine_mg shows that while most people stay within the "Safe" range (centered around the mean of 238mg), there is a significant "right tail" of high-consumers. This visual spread justifies the need to split the data into "High" and "Safe" groups to see if that tail end has different health outcomes.

Sample Consistency: The Age distribution shows a relatively uniform spread across the 18-80 range, ensuring that conclusions aren't biased toward just young or just elderly participants.

3.2.2 Categorical Variables Visualization

Frequency distributions for categorical variables were evaluated using normalized horizontal bar charts. Calculating the percentage distribution of categories (e.g., Gender, Occupation, and Health Issues) ensures that our subsequent inferential models are based on a well-represented and balanced sample population.



Group Imbalance: The most critical takeaway is the Caffeine_Group split. Since approximately 87% of people are in the "Safe" group and only 13% are in the "High" group, a standard T-test might be inaccurate, which suggests the use of the Welch T-test as it is specifically designed to handle groups of unequal sizes. Lifestyle Context: The Smoking and Alcohol_Consumption charts show that the majority of the population are non-smokers and non-drinkers. Diversity: With 20 countries and 5 occupations represented almost equally, we can conclude that the dataset is diverse.

4 Statistical Methods

We employ two main statistical methods to analyze the dataset:

4.1 Hypothesis Testing (T-test)

To evaluate the 400mg threshold, the population was partitioned into two cohorts: the "Safe" group ($\leq$ 400mg/day) and the "High" group ($>$ 400mg/day). We performed a Welch Two-Sample t-test to compare mean resting heart rates between these groups.The Welch t-test was selected specifically for its robustness against unequal variances (heteroscedasticity) and unequal sample sizes. High caffeine intake is often associated with higher variance in physiological responses; thus, the Welch method provides a more reliable p-value than the standard Student's t-test by adjusting the degrees of freedom via the Satterthwaite equation: The test statistic t is calculated using the Welch-Satterthwaite equation:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups.
- s_1^2 and s_2^2 are the sample variances.
- n_1 and n_2 are the sample sizes.

Hypotheses:

- $H_0 : \mu_{High} = \mu_{Safe}$ (The mean heart rate is equal across both groups).
- $H_1 : \mu_{High} > \mu_{Safe}$ (The mean heart rate of the high-consumption group is significantly higher).

4.2 Multiple Linear Regression

To assess the simultaneous impact of multiple lifestyle factors, we employed Multiple Linear Regression (MLR). This allows us to isolate the effect of caffeine while controlling for potential confounders such as sleep, physical activity, and BMI. The model is defined by the following equation:

The theoretical model for our Multiple Linear Regression is defined as:

$$Y_{HR} = \beta_0 + \beta_1 X_{Caffeine} + \beta_2 X_{Sleep} + \beta_3 X_{BMI} + \beta_4 X_{Smoking} + \beta_5 X_{Activity} + \epsilon$$

Where:

- Y_{HR} is the dependent variable (Resting Heart Rate).
- β_0 is the y-intercept (the value of Y when all X are zero).
- $\beta_1, \ldots, \beta_5$ are the partial regression coefficients for each predictor.
- ϵ is the error term (residuals).

Where β_0 represents the intercept, $\beta_{1\ldots5}$ represent the partial regression coefficients for each predictor, and ϵ represents the error term. We assessed the significance of individual predictors using t-tests for each coefficient and evaluated the global model fit using the F-statistic and Adjusted R^2.

5 Results

We seperate the data into <=400mg and >400mg of caffeine intake and create a new column called Caffeine_Group to store this information.

This is how many people are in each group:

```
##
## High (>400mg) Safe (<400mg)
##          1246          8754
```

Then we run our two sampe t test to compare the mean heart rates of the two groups: and this is the result of it

```
##
##  Welch Two Sample t-test
##
## data:  Heart_Rate by Caffeine_Group
## t = 3.3169, df = 1619.3, p-value = 0.0009305
## alternative hypothesis: true difference in means between group High (>400mg) and group Safe (<400mg)
## 95 percent confidence interval:
##  0.4032072 1.5701557
## sample estimates:
## mean in group High (>400mg) mean in group Safe (<400mg)
##                    71.48154                    70.49486
```

The results yield a p-value of 0.00093, which is significantly lower than the standard alpha level of 0.05. This allows us to reject the null hypothesis (H_0). We observe that the "High" consumption group has a mean heart rate of 71.48 bpm, while the "Safe" group averages 70.49 bpm. The 95% confidence interval [0.40, 1.57] indicates that exceeding the 400mg limit is associated with a consistent increase in resting heart rate.

5.0.1 Multivariate Predictors of Heart Rate

To understand which lifestyle factors most influence heart rate, we ran a Multiple Linear Regression model.

```
##
## Call:
## lm(formula = Heart_Rate ~ Caffeine_mg + Sleep_Hours + BMI + Smoking +
##     Physical_Activity_Hours, data = coffee_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.567  -6.833  -0.066   6.638  36.917
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             71.6123306  0.8719005  82.134  < 2e-16 ***
## Caffeine_mg              0.0039216  0.0007251   5.408 6.51e-08 ***
## Sleep_Hours             -0.2070377  0.0817360  -2.533   0.0113 *
## BMI                     -0.0220116  0.0251029  -0.877   0.3806
## SmokingYes               0.1534319  0.2449773   0.626   0.5311
## Physical_Activity_Hours -0.0077872  0.0227230  -0.343   0.7318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.804 on 9994 degrees of freedom
## Multiple R-squared:  0.004367,   Adjusted R-squared:  0.003869
## F-statistic: 8.767 on 5 and 9994 DF,  p-value: 2.611e-08
```

The model is statistically significant ($p < 0.001$), identifying two primary drivers of heart rate:

- Caffeine Intake ($p < 0.001$): Every milligram of caffeine increases the heart rate by approximately 0.0039 bpm. This confirms caffeine's role as a physiological stimulant.
- Sleep Hours ($p = 0.011$): Sleep has a significant negative coefficient (-0.207), suggesting that for every additional hour of sleep, the heart rate decreases.

Surprisingly, variables such as Smoking, BMI, and Physical Activity did not reach statistical significance in this specific model, suggesting their linear impact is secondary to the immediate effects of caffeine and sleep.

6 Conclusion

Based on the statistical analysis of 10,000 individuals, we can draw three primary conclusions regarding the relationship between caffeine consumption and cardiovascular health:

1. The 400mg Threshold is Valid: Our hypothesis testing confirms that consuming more than 400mg of caffeine per day leads to a statistically significant increase in resting heart rate. This provides empirical support for the guidelines issued by health organizations like the FDA.
2. Sleep as a Modulator: The regression model highlights that sleep is a critical protective factor. While caffeine stimulates the heart, adequate sleep works to lower the resting heart rate, highlighting the importance of "sleep hygiene" for regular coffee drinkers.
3. Statistical Significance vs. Practical Magnitude: While the results are highly significant, the low Adjusted R^2 (0.0038) indicates that lifestyle factors explain only a small fraction of the total variation in heart rate. This suggests that while caffeine and sleep matter, biological and genetic factors likely play a much larger role in determining an individual's baseline heart rate.