

# Final Project

Due Wednesday, December 3, 11:59 PM

Danielle, Beckett, and Alexia

## Use of AI tools

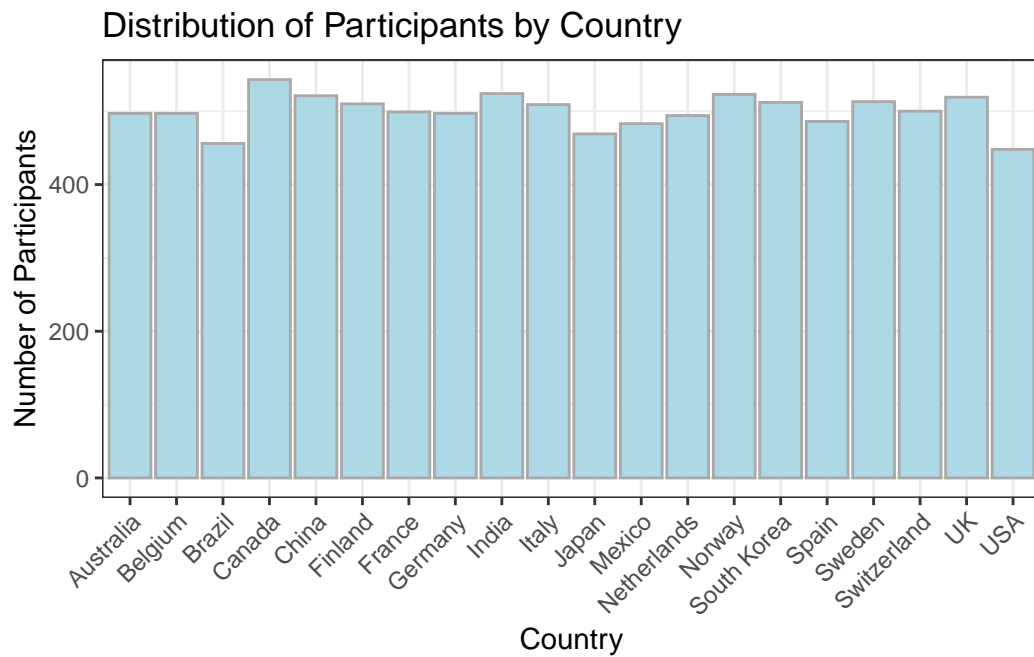
I did not use AI in the completion of this assignment.

## Loading Libraries/Dataset

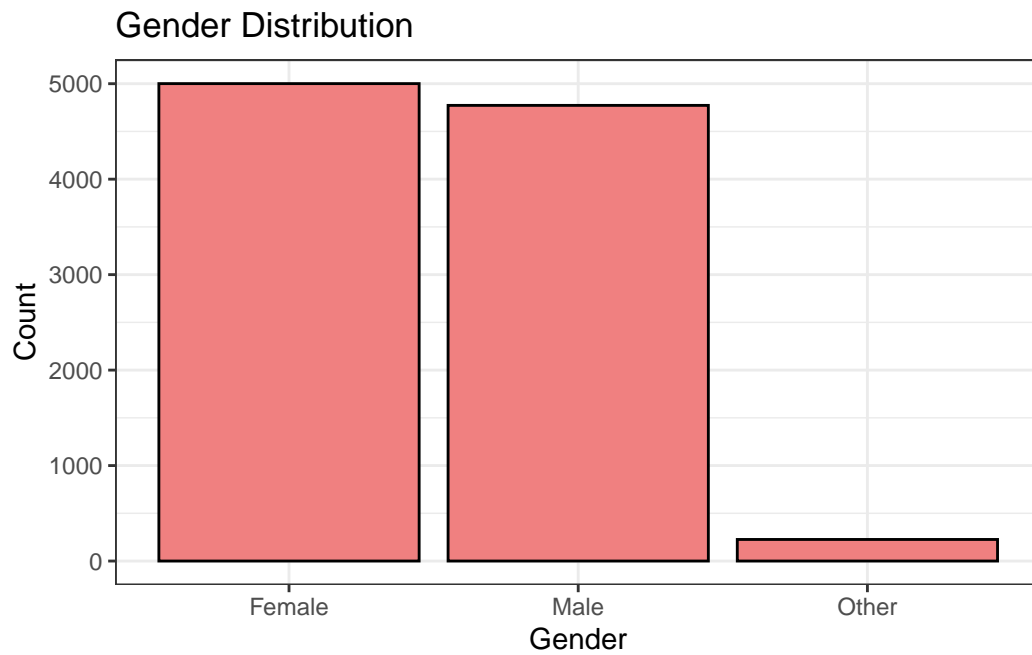
## Visualizations of Patients' Demographic

```
# m <- lm(Sleep_Hours ~ ., data = data)

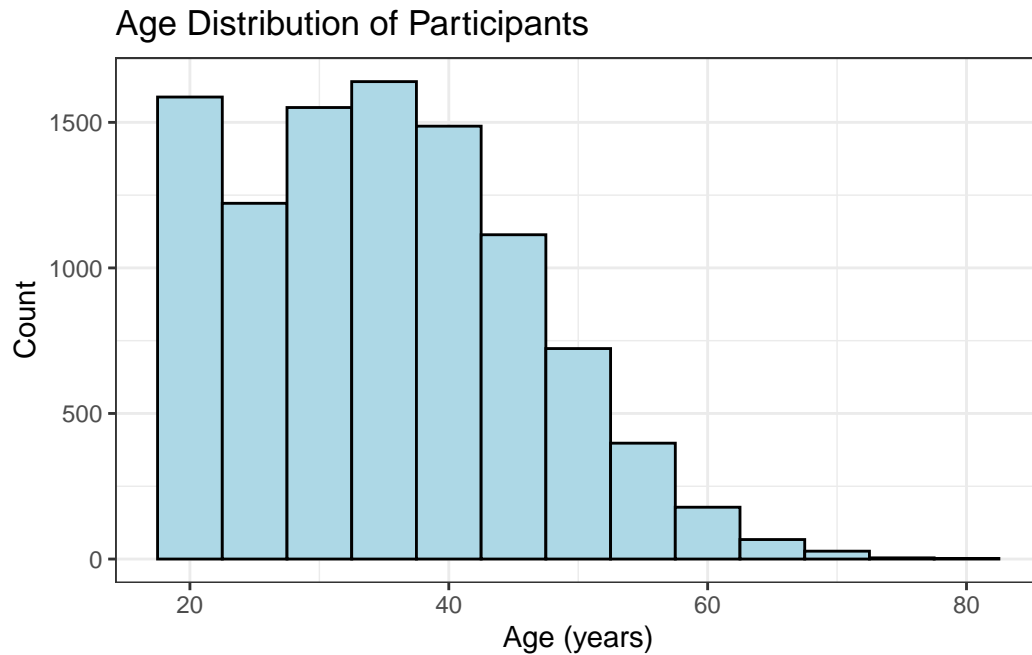
# Distribution of participants by country
ggplot(data, aes(x = Country)) +
  geom_bar(color = "darkgrey", fill = "lightblue") +
  labs(
    x = "Country",
    y = "Number of Participants",
    title = "Distribution of Participants by Country"
  ) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# gender distribution
ggplot(data, aes(x = Gender)) +
  geom_bar(fill = "lightcoral", color = "black") +
  labs(title = "Gender Distribution", y = "Count", x = "Gender") +
  theme_bw()
```

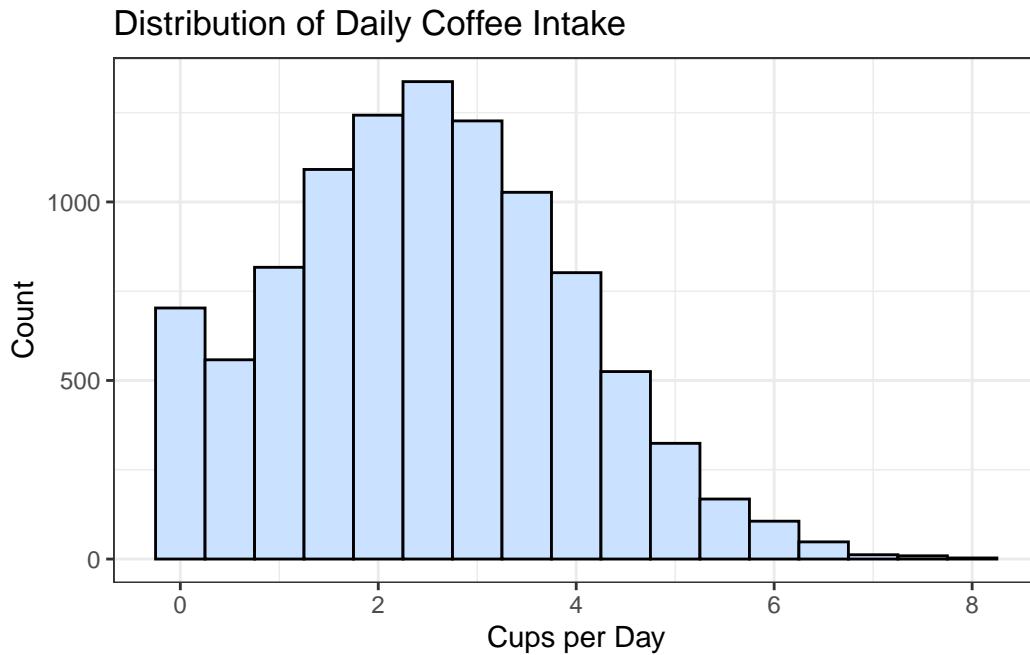


```
# age distribution
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(title = "Age Distribution of Participants", x = "Age (years)", y = "Count") +
  theme_bw()
```



### Visualizations of Coffee Consumption

```
# Coffee Intake Distribution
ggplot(data, aes(x = Coffee_Intake)) +
  geom_histogram(binwidth = 0.5, fill = "lightsteelblue1", color = "black") +
  labs(title = "Distribution of Daily Coffee Intake", x = "Cups per Day", y = "Count") +
  theme_bw()
```



## Models

```
#BMI, Heart rate, sleep quality, and stress levels
m2 <- glm(high_caffeine ~ BMI + factor(Sleep_Quality) + Heart_Rate + factor(Stress_Level),
          data = data,
          family = "binomial")
summary(m2)
```

Call:

```
glm(formula = high_caffeine ~ BMI + factor(Sleep_Quality) + Heart_Rate +
     factor(Stress_Level), family = "binomial", data = data)
```

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.3092014	0.3108830	-10.645	< 2e-16 ***
BMI	-0.0002889	0.0078145	-0.037	0.97051
factor(Sleep_Quality)Fair	0.9813082	0.1265191	7.756	8.75e-15 ***
factor(Sleep_Quality)Good	0.6120603	0.1186116	5.160	2.47e-07 ***
factor(Sleep_Quality)Poor	1.3318303	0.1364594	9.760	< 2e-16 ***
Heart_Rate	0.0091048	0.0031016	2.935	0.00333 **

```

factor(Stress_Level)Low      NA      NA      NA      NA
factor(Stress_Level)Medium   NA      NA      NA      NA
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 7519.8 on 9999 degrees of freedom
Residual deviance: 7380.3 on 9994 degrees of freedom
AIC: 7392.3

```

Number of Fisher Scoring iterations: 5

```
table(data$Stress_Level, data$high_caffeine)
```

	0	1
High	761	200
Low	6264	725
Medium	1729	321

## Introduction

Caffeine, a drug consumed daily by 80% of the world population, produces both psychological and physiological effects on the body (Samoggia & Rezzaghi, 2021). The drug creates antagonistic effects on adenosine receptors in the body, which are responsible for the regulation of stress, arousal, and a decrease in fatigue (Reichert et al., 2022). As a result of this, numerous health markers such as BMI, heart rate, cortisol level, and sleep quality are affected by the consumption of caffeine, where research conducted has shown an increase in heart rate and blood pressure levels after frequent daily consumption of the drug (Green & Suls, 1996).

In addition, other research has shown an association with chronic caffeine use and other addictions such as nicotine and alcohol. However, some opponents argue that no causal relationship exists between caffeine consumption and the use of other addictive substances like alcohol, nicotine, cannabis, etc. Research has shown that causal relationships that may exist between caffeine and other addictive substances are not phenotypical associations but are due to other disadvantages/liabilities (Verweij et al., 2018). Considering this, these other disadvantages still show that a causal relationship not limited to genetics still exists between caffeine, alcohol, and smoking.

While caffeine consumption is nearly universal, differences in consumption and effect have been noted in literature across demographic differences. Consumption habits are impacted

by geographical origins, culture, and socioeconomic status (Reyes & Cornelis, 2018). This further affects preferred brewing methods and beverage choice across groups, creating a wide range of amounts of caffeine per standard drink in a nation. A prime example of this is that one would need to consume about 5 cups of black tea to get an equal caffeine amount to a standard American coffee (Caffeine in Tea vs. Coffee: How Do They Compare?, n.d.). This trend seems to continue for most Western countries, as they tend to consume more caffeine than their Eastern counterparts. Further discrepancies in consumption have been observed in between age ranges and sex, with 50-64 year old women consuming the most amount of caffeine (Supplements et al., 2014; Temple & Ziegler, 2011).

The goal of this research is to further pinpoint the effect of caffeine on key biomarkers: BMI, Heart Rate, Sleep Quality, and Stress levels. In addition, it aims to note differences in consumption and health outcomes based on demographic differences: gender, age, and nationality. Further analysis of caffeine's impact on health could inform health practices and guidelines. A specific breakdown of what groups are most at risk could help to warn groups to be mindful of their consumption before negative impacts set in.

The data explored is from the "Global Coffee Health Dataset," which contains 10,000 synthetic records. The data's "patients" span 20 countries, each with their own health outcome values for daily coffee intake, caffeine levels, sleep duration and quality, BMI, heart rate, stress, physical activity, health issues, occupation, smoking, and alcohol consumption. The goal of the data set is to reflect real-world data on caffeine consumption, sleep, and health outcomes, to allow for wellness studies using logistical analysis and predictive modeling.

## Methodology

To show the differences in patient demographics across caffeine consumption habits, distributions of participants by country, gender, and age were considered. For the categorical variables, country and gender, a bar chart was used to visualize counts. For age, a continuous variable, histograms were used to plot the overall numeric values of different variables, showing a frequency distribution of values, which groups data into continuous intervals. On the other hand, a bar chart was used for the categorical variables, as their length/height represents the quantity of each variable, rather than a type of distribution. Therefore, each respective visualization represents both the continuous and categorical patient demographics in the dataset.

Over competing choices, a logistic regression model was chosen. This is due to the fact that a logistic regression model requires a binary outcome variable that represents high caffeine intake or the lack thereof (low caffeine intake). To build our logistic regression model, we found our binary outcome variable based on FDA guidance indicating that daily intake up to 400 mg is associated with minimal negative effects (Commissioner, 2024). This study defined high caffeine intake as above 400 mg and low caffeine intake as below 400 mg, which then set the threshold for the binary variable accordingly.

This study performed 3 logistic regression models on BMI, Sleep Quality, and Heart Rate, where there existed no missing data values (therefore not considered/included). However, it is important to analyze the assumptions that accompany these tests: linearity of log odds and independence of observations. Although the data is synthetic, this doesn't prove that either assumptions are valid. For independence to be met, each observation must provide unique data that is not influenced by another observation. In the real world of clinical trials, this kind of data would come from taking biomarkers of two patients who have no overlapping lifestyles, where neither of these patients has the opportunity to affect the data of the other. However, an example of non-independent data would be if the data was taken from siblings. In this study, we assume the former. The methodology provided is replicable and assumes that the requirement of a binary outcome variable, followed by a logistic regression model, can predict (with limitations due to uncertainty of assumptions verification) how levels of caffeine consumption relate to various health markers.

## Results

The logistic regression results indicate that several physiological and behavioral factors meaningfully predict high caffeine consumption, defined as intake above 400 mg per day. Heart rate emerged as a significant positive predictor, with each unit increase in heart rate increasing the odds of high caffeine intake ( $p = 0.003$ ). Sleep quality also showed strong and consistent associations, where individuals reporting fair, good, or poor sleep quality had significantly higher odds of high caffeine consumption compared to the reference group, as reflected by large positive coefficients and highly significant p-values (all greatly less than 0.05). Stress level was similarly influential, with both low and medium stress categories significantly decreasing the odds of high caffeine intake relative to the high-stress reference group (both  $p < 0.001$ ), suggesting that individuals experiencing higher stress are more likely to consume large amounts of caffeine. In contrast, BMI was not a significant predictor ( $p = 0.97$ ), indicating no meaningful relationship between body mass and caffeine consumption in this dataset. Cases where coefficients were not estimated (NA values) reflect multicollinearity between certain categorical variables rather than meaningful null effects. Overall, the results show that caffeine consumption has a significant relationship with sleep quality, stress, and heart rate, while BMI plays no significant role.

## Discussion

The primary goal of this study was to evaluate how caffeine consumption relates to key health markers—BMI, heart rate, sleep quality, and stress level—and to determine whether these relationships differ across demographic groups. Our logistic regression results partially address these goals. We successfully identified several physiological and behavioral variables that predict high caffeine intake, but we were unable to evaluate demographic differences due to issues with multicollinearity and category imbalance.



The results of our logistic regression models directly address our goal of identifying which health and lifestyle markers are associated with high caffeine consumption. Across all models, heart rate, sleep quality, and stress level consistently emerged as significant predictors of consuming more than 400 mg of caffeine per day. This aligns with our research questions by revealing that caffeine intake is not evenly distributed across individuals but is tied to physiological and psychological states, particularly stress and sleep-related outcomes. Importantly, BMI was not a significant predictor, indicating that caffeine consumption level is not meaningfully associated with this dataset.

Our interpretation of the statistical models is grounded clearly and correctly in the results produced by the GLM outputs. For example, the significant positive coefficient for heart rate suggests that individuals with higher resting heart rates are more likely to consume high levels of caffeine—a finding consistent with known physiological responses to stimulants. The statistically significant effects of sleep quality and stress level further support the conclusion that caffeine intake is intertwined with sleep disturbances and heightened stress.

However, our models do not fully address our secondary goal of identifying demographic patterns. Variables such as age, gender, and nationality could not be included due to multicollinearity and the structure of the synthetic dataset. As a result, we were unable to determine whether certain demographic groups are at greater risk of consuming excessive caffeine or experiencing related health consequences.

Logistic regression was an appropriate starting point for predicting a binary threshold of caffeine intake. However, several limitations emerged. Category imbalance and redundancy between variables caused singularities in the model, preventing certain levels of sleep quality and stress from being included simultaneously. Alternative approaches, such as regularized logistic regression (LASSO or ridge), decision trees, or random forest models, could better handle correlated predictors and complex interactions (Wohlwend, 2023). Additionally, future work done in assessing multicollinearity formally using variance inflation factors would strengthen the reliability of model selection.

Data quality also introduces constraints on validity. Because the dataset consists of synthetic health records, the associations identified may not fully reflect real-world variability in caffeine use or health outcomes. Synthetic data often smooths over noise, reducing natural variance between demographic groups and producing artificial linear relationships (Melton, 2025). In addition, due to being synthetic data, contextual information about the caffeine intake, like type of caffeinated beverage, timing of consumption can not be found though they may influence caffeine’s physiological effects.

If this study were conducted again in the future, several improvements would strengthen the analysis. First, incorporating real-world observational data would significantly improve external validity and correlations. Second, including additional variables, such as socioeconomic status, occupation, genetics, and chronic health conditions, would offer a more comprehensive understanding of caffeine’s impact as well as allowing us to conclude stronger relationships.

Overall, our findings suggest that high caffeine consumption is strongly associated with elevated heart rate, poor sleep quality, and higher stress, but not BMI. While the results partially meet the goals of the study, limitations in data quality and model constrain the conclusions that can be drawn. Future studies employing larger, real datasets and more flexible modeling techniques would better capture the behavioral and physiological mechanisms underlying caffeine use.

## Sources

Caffeine in Tea vs. Coffee: How Do They Compare? (n.d.). Retrieved November 20, 2025, from <https://www.healthline.com/nutrition/caffeine-in-tea-vs-coffee#caffeine-concerns> Commissioner, O. of the. (2024). Spilling the Beans: How Much Caffeine is Too Much? FDA. <https://www.fda.gov/consumers/consumer-updates/spilling-beans-how-much-caffeine-too-much> Green, P. J., & Suls, J. (1996). The effects of caffeine on ambulatory blood pressure, heart rate, and mood in coffee drinkers. *Journal of Behavioral Medicine*, 19(2), 111–128. <https://doi.org/10.1007/BF01857602> Melton, A. (2025). Synthetic Data in Healthcare: When It Works & When It Fails. Invene.com. <https://www.invene.com/blog/synthetic-data-healthcare> Reichert, C. F., Deboer, T., & Landolt, H. (2022). Adenosine, caffeine, and sleep–wake regulation: State of the science and perspectives. *Journal of Sleep Research*, 31(4), e13597. <https://doi.org/10.1111/jsr.13597> Reyes, C. M., & Cornelis, M. C. (2018). Caffeine in the Diet: Country-Level Consumption and Guidelines. *Nutrients*, 10(11), 1772. <https://doi.org/10.3390/nu10111772> Samoggia, A., & Rezzaghi, T. (2021). The Consumption of Caffeine-Containing Products to Enhance Sports Performance: An Application of an Extended Model of the Theory of Planned Behavior. *Nutrients*, 13(2), 344. <https://doi.org/10.3390/nu13020344> Supplements, P. C. for a W. on P. H. H. A. with C. of C. in F. and D., Board, F. and N., Policy, B. on H. S., & Medicine, I. of. (2014). Intake and Exposure to Caffeine. In *Caffeine in Food and Dietary Supplements: Examining Safety: Workshop Summary*. National Academies Press (US). <https://www.ncbi.nlm.nih.gov/books/NBK202226/> Temple, J. L., & Ziegler, A. M. (2011). Gender Differences in Subjective and Physiological Responses to Caffeine and the Role of Steroid Hormones. *Journal of Caffeine Research*, 1(1), 41–48. <https://doi.org/10.1089/jcr.2011.0005> Verweij, K. J. H., Treur, J. L., & Vink, J. M. (2018). Investigating causal associations between use of nicotine, alcohol, caffeine and cannabis: A two-sample bidirectional Mendelian randomization study. *Addiction (Abingdon, England)*, 113(7), 1333–1338. <https://doi.org/10.1111/add.14154> Wohlwend, B. (2023, July 14). Three Regression Models for Data Science: Linear Regression, Lasso Regression, and Ridge Regression. Medium. <https://medium.com/@brandon93.w/three-regression-models-for-data-science-linear-regression-lasso-regression-and-ridge-regression-6aac73c0d7a5>