Coffee, Health, and Lifestyle Analysis

Author: Damian Grochowski

Tools: Python, Pandas, Matplotlib/Seaborn, Jupyter Notebook, Scikit-learn

Overview:

The purpose of this project is to explore the relationship between coffee/caffeine consumption and various health indicators using realistic correlations observed in research. The goal is to analyze trends, identify correlations, and visualize how coffee intake might be associated with outcomes such as sleep, stress, and health, while also building a predictive model for these outcomes. The sub-goal of this project is to highlight the challenges of analyzing synthetic yet realistic datasets, including data leakage and inaccurate predictions.

Dataset:

Source: https://www.kaggle.com/datasets/uom190346a/global-coffee-health-dataset
Size: 10,000 x 16
Key Variables:

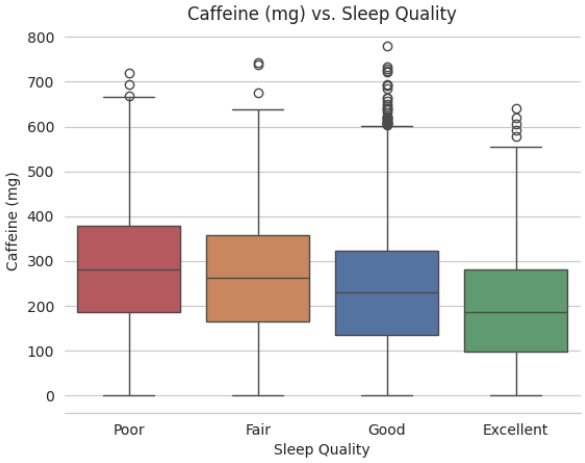
Coffee_Intake	Float	Daily coffee consumption in cups (0–10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)

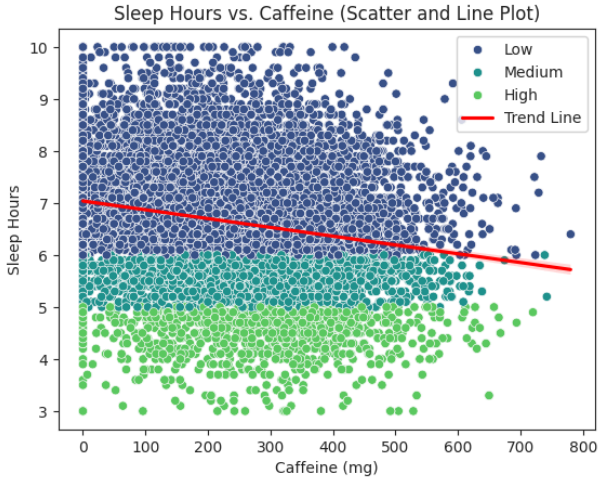
Data Cleaning & Preparation:

One of the first things I did after uploading the dataset was print a summary of it with the .info() function. The summary showed no null values except for in one column: “Health_Issues”, which had 5941 null values. Upon closer inspection of these null values, I found that the null values represented a person with no health issues. After this realization, I replaced every null value in Health_Issues with the string “None”. Besides this, I saw nothing else out of the ordinary that might have required cleaning. The next thing I did was engineer a new feature into my dataset called “Caffeine_Dose”. This feature breaks down whether someone drinks a low (less than 200 mg), moderate (between 200 mg and 400 mg), or high (more than 400 mg) amount of caffeine based on FDA recommendations.

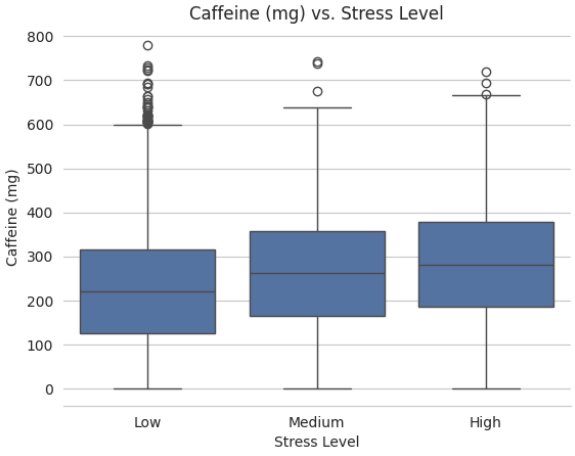
Exploratory Data Analysis (EDA):

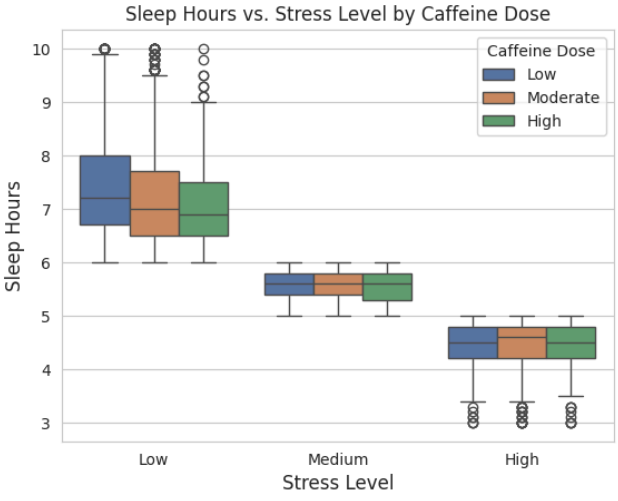
After cleaning and preparing my data, it was time to start analyzing it through visualizations. The first thing I did was create a histogram of all my numerical data just to gain an understanding of some quick correlations. After noticing nothing there, I started my analysis with the effects of coffee/caffeine on sleep and stress.





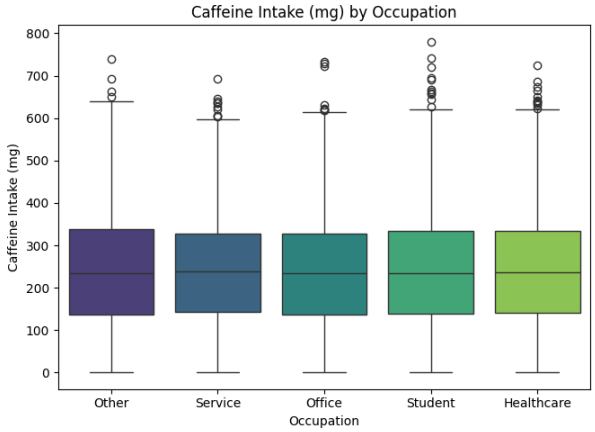
There was actually a noticeable difference between the average amount of caffeine a person consumed and their quality of sleep. The boxplot shows that a person with poor sleep consumes a median caffeine amount of almost 300 mg, while a person with excellent sleep has a median of almost 200 mg, showing a negative correlation between caffeine consumption and sleep quality. This is also shown in the scatter plot to the right, which displays the amount of sleep hours decreasing as caffeine consumption increases.

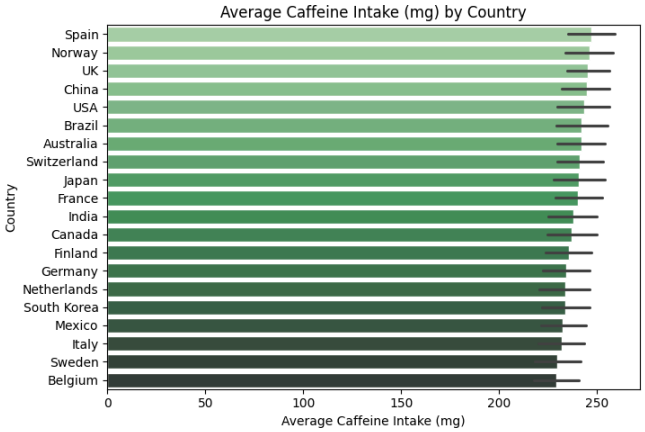




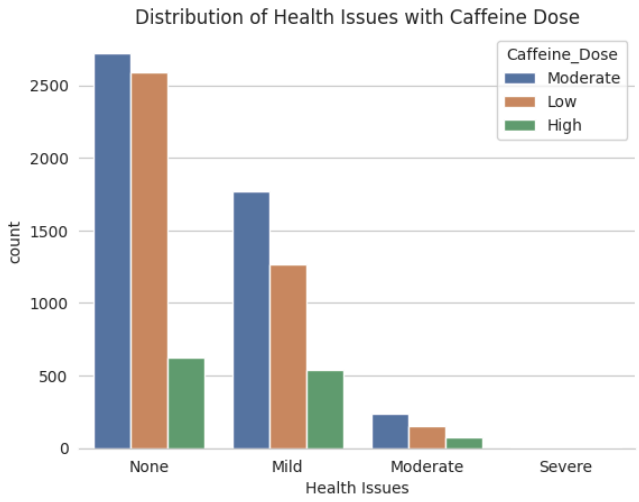
For stress level, we can see some increase in stress as caffeine consumption increases. With low stress having a median of close to 225 mg of caffeine, and high stress having a median of close to 275 mg. However, due to the synthetic nature of the dataset, stress is fully based on sleep hours, as shown in the box graph to the right. This will make it much harder to predict both

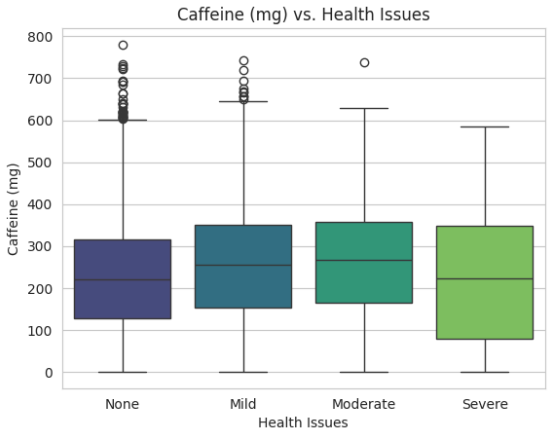
sleep_quality and stress_level when there is such a large correlation between them. Despite this, caffeine still has some effect, as shown by the low stress level column on the graph to the right.



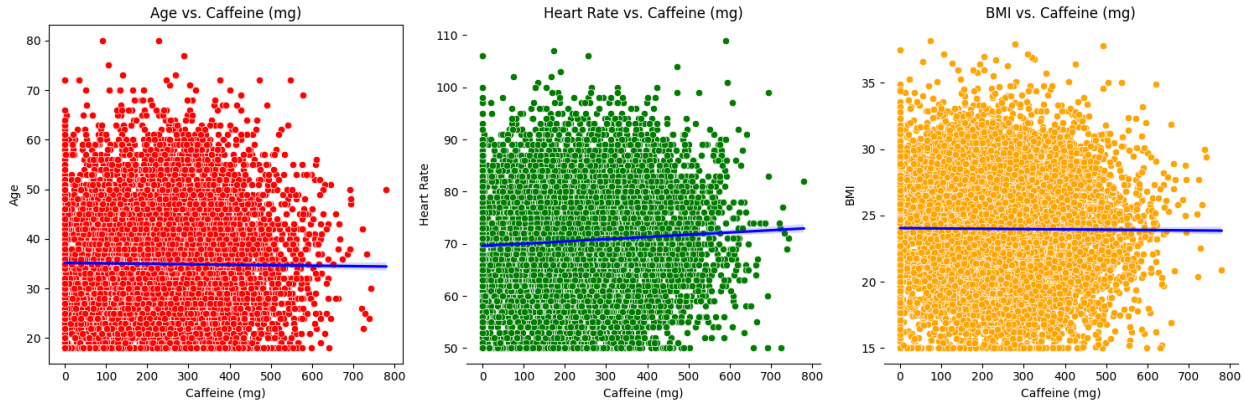


Next, I analyzed some lifestyle factors related to caffeine consumption. As shown by the graphs above, caffeine intake is not really affected by features such as occupation and country. The distribution of caffeine intake is relatively stagnant when comparing each occupation. While the caffeine intake average is highest in Spain with 247 mg, and lowest in Belgium with 229.08 mg.





As shown by these graphs, theres no apparent correlation between caffeine intake and health issues.



Caffeine also little to no effect on other health attributes like age, heart rate (which has a slight increase as caffeine increases), and BMI.

Sleep Quality and Stress Level Prediction Model:

One of the biggest problems with this dataset is that the features I am predicting (Sleep_Quality and Stress_Level) are directly based on other features of the dataset. This causes a serious case of data leakage which abnormally high and inaccurate predictions. However, I decided to take this moment as a learning opportunity to showcase the problem of data leakage and create a predictive model that shimmies around said problem.

Model Preparation:

To prepare the model I had to split up the data between my target variable and features. For sleep quality and stress level with data leakage, I used all features for my prediction, excluding the target variable. For sleep quality without data leakage I dropped any variables that are related to sleep_quality or are redundant, these included Sleep_Hours, Stress_Level, Health_Issues (As Health_Issues is based on Sleep_Quality), etc. For stress level I did the same. Then I split the data by numerical and categorical variables and created a column transformer which passes through the numerical features and one hot encodes the categorical features. I then split the data into training (X_Train) and testing (X_test) data. I then applied the column transformer onto my training and test data.

Model Predictions:

I trained and tested three different models to predict, these were a Random Forest Classifier, K-Nearest-Neighbor Classifier, and Logistic Regression. The baseline accuracy for sleep quality is 56.40%, which is from its most frequent class “Good”. The baseline accuracy for stress level is 69.90% which is from its most frequent class “low”. For sleep quality with leakage, the highest accuracy was 98.90% from Logistic Regression. For stress level with leakage, the highest accuracy was 100% from the Random Forest Classifier. For sleep quality without leakage, the highest accuracy was 56.65% from the Random Forest Classifier. For stress level without leakage, the highest accuracy was 69.90% from KNN.

Interpreting the Model Results:

The models were able to predict incredibly (even abnormally) well the classes of sleep quality and stress levels when they included data leakage. However, the models performed extremely poor (never reaching above the baseline) for the data without the leakage. The confusion matrix for sleep quality was extremely biased towards the “Good” class with almost every variable being predicted as being in that class and the same can be said toward stress level with the “Low” class. The most important features for the predictions with leakage included Stress_Level, Sleep_Hours and Health_Issues for sleep quality and, Sleep_Hours and Sleep_Quality for stress level. Without this leakage, we can see that the most important features for both variables were Caffeine_mg at the top, followed by BMI, Physical_Activty_Hours, and Heart_Rate.

Conclusion:

It is clear from my analysis that this dataset cannot be used to make meaningful predictions to its variables. From its synthetic and assumptive natures to its data leakage caused by creating variables based off its own variables, this cannot reliably predict health features such as sleep quality and stress levels. Despite this, I feel like it is still a very interesting topic that provides some fascinating insights, especially in terms of the effects of coffee. Without the data leakage features, caffeine became the most important feature for predicting sleep quality and stress levels. The visual analysis also showed this trend, with a negative correlation between caffeine and said variables. Caffeine can also be shown to have some effects of health features like heart rate. Even though we can’t predict with a lot of accuracy, it can be induced that drinking coffee has effects on a persons health. Creating or find a dataset or study that is more accurate about these variables should be the next steps to truly figure out coffee and its correlations.