Coffee Health Analysis

Table of Contents

Objective

To explore if caffeine affects sleep quality/time.

Data Source

Sourcing

Data comes from Kaggle and can be found here:

https://www.kaggle.com/datasets/uom190346a/global-coffee-health-dataset

Collection

Records are synthetic (made-up), but designed around real-world data.

Contents

The columns are as follows:

- ID
- Age
- Gender
- Country
- Coffee Intake
- Caffeine mg
- Sleep Hours
- Sleep Quality
- BMI
- Heart Rate
- Stress Level
- Physical Activity Hours
- Health Issues
- Smoking
- Alcohol Consumption

Relevance

While the data is synthetic, the fact that it’s designed to mimic real data is still useful (though, it would be nice to know how that was achieved).

Limitations

Obviously, the data isn’t real so it can’t be used to extrapolate perfectly. Similarly, there’s no need to worry about collection problems because the data was created instead of collected. However, bias could still be present depending on the model the data was created upon.

There’s also a lack of dates in the data so we can’t make any inferences over time.

Data Profile

Cleaning Data

Min

Mean

Mode

Max

Outliers

Frequency

Duplicates

Missing Values

Understanding Data

Variables and Data Types:

Column Name	Qualitative or Quantitative?	Time Variance	Data Type
ID	Quantitative	time-invariant	discrete
Age	Quantitative	time-variant	discrete
Gender	Qualitative	time-invariant	nominal
Country	Qualitative	time-invariant	nominal
Coffee Intake	Quantitative	time-variant	continuous
Caffeine mg	Quantitative	time-variant	continuous
Sleep Hours	Quantitative	time-variant	continuous
Sleep Quality	Qualitative	time-variant	ordinal
BMI	Quantitative	time-variant	continuous
Heart Rate	Quantitative	time-variant	discrete
Stress Level	Qualitative	time-variant	ordinal
Physical Activity Hours	Quantitative	time-variant	continuous
Health Issues	Qualitative	time-variant	nominal
Smoking	Quantitative	time-invariant	discrete (binary)
Alcohol Consumption	Quantitative	time-invariant	discrete (binary)

Data Integrity Issues:

- There are 5,941 null records in health issues. However, these records are showing participants as having no health issues and should not be removed.

Changed/Fixed Records:

- Changed null values to say "none"
- Changed data types for smoking and alcohol consumption to be more clear (binary/Boolean)

Summary:

- 16 Variables, 10,000 records

Qualitative:

- Gender

Female	5001
Male	4773
Other	226

- Country

Canada	543
India	524
Norway	523
China	521
UK	519
Sweden	513
South Korea	512
Finland	510
Italy	509
Switzerland	500
France	499
Germany	497

Australia	497
Belgium	497
Netherlands	494
Spain	486
Mexico	483
Japan	469
Brazil	456
USA	448

- Sleep Quality

Poor	961
Fair	2050
Good	5637
Excellent	1352

- Health Issues

None	5941
Mild	3579
Moderate	463
Severe	17

Quantitative:

- Age (might be worth a histogram)
  - Count: 10,000
  - Minimum: 18
  - Maximum: 80
  - Mean: 35
  - Median: 34
  - Mode: 18
- Coffee Intake (might be worth a histogram)
  - Count: 10,000
  - Minimum: 0
  - Maximum: 8.2
  - Mean: 2.5
  - Median: 2.5
  - Mode: 0 (558 of them! That’s suspicious, maybe a control group?)
- Caffeine mg (might be worth a histogram)
  - Count: 10,000
  - Minimum: 0
  - Maximum: 780.3
  - Mean: 238.4
  - Median: 235.4
  - Mode: 0 (528 counts. Generally confirms that caffeine and coffee are linked, but makes me wonder what happened to the 30 people that had coffee but no caffeine. Might be worth exploring for further cleaning/wrangling).

- Sleep Hours (might be worth a histogram)
  - Count: 10,000
  - Minimum: 3
  - Maximum: 10
  - Mean: 6.6
  - Median: 6.6
  - Mode: 6.7
- BMI (might be worth a histogram)
  - Count: 10,000
  - Minimum: 15
  - Maximum: 38.2
  - Mean: 23.98
  - Median: 24
  - Mode: 15 (However, the next most popular values are from 23.3-25.5
- Heart Rate (might be worth a histogram)
  - Count: 10,000
  - Minimum: 50
  - Maximum: 109
  - Mean: 70.6
  - Median: 71
  - Mode: 70

- Physical Activity Hours (might be worth a histogram)
  - Count: 10,000
  - Minimum: 0
  - Maximum: 15
  - Mean: 7.48
  - Median: 7.5
  - Mode: 8.4

Ethics and Limitations

There is no personal data and the rest is fictitious. As mentioned earlier, the model the data was created upon could've had biases that would affect the integrity of this data. However, since coffee is pretty well studied, I tend to think there's little left glean (though studies about the benefits of decaf coffee and how they compare to caffeinated coffee would be interesting).

Questions to explore

Why/how does coffee impact sleep?

Does coffee offer benefits that make up for the lack of good sleep quality/time?

How important is the timing of coffee consumption? Can you time it to not affect sleep?

Are certain ages more likely to drink coffee?

How about different genders?

How important is the region to coffee consumption? Is that a cultural bias or a productivity bias?

Do different types of coffee produce different effects? (Black, mocha, latte, etc).

What differences can be found from the region the beans came from?

Does temperature matter?

Extraction method?

Grinder type or grind size?

Filter types?

Age of beans or grounds?

Type of drinking vessel (glass, ceramic, vacuum/double-walled bottle, etc)?

Additives? (Sugar, cream, flavorings, etc).

Are these people getting their coffee from home or a café of some sort?