**DSA210**

**EFFECT OF DIFFERENT VARIABLES ON SLEEP QUALİTY**

**ERTUĞRUL GÖR**

**29067**

1. **Intrduction**

This project investigates how daily habits affect sleep quality by analyzing a self-recorded dataset. Variables such as exercise, caffeine intake, screen time, sugar and carbohydrate consumption, and sleep duration were examined to understand their influence on deep sleep and sleep efficiency. After creating and enriching the dataset, we conducted exploratory data analysis (EDA), hypothesis testing, and time-series visualization. Finally, we applied a Random Forest Regressor model to predict the Sleep Quality Score using the most relevant lifestyle features. The results showed that screen time, caffeine, and exercise levels significantly impact sleep quality.

**Parameters in the Report**

The dataset used in this study includes the following key parameters, collected daily over a two-month period:

* **Date**: The record date of the observation.
* **Exercise (Fitness) (min)**: Duration of structured physical activity such as gym or cardio training.
* **Exercise (Walking) (steps)**: Number of steps taken during the day.
* **Calories Burned**: Estimated total calories burned, calculated based on physical activity.
* **Carbohydrates (g)**: Daily carbohydrate intake in grams.
* **Sugar (g)**: Daily sugar intake in grams.
* **Protein (g)**: Daily protein intake in grams.
* **Total Calories**: Total daily caloric intake from food.
* **Caffeine (mg)**: Daily caffeine consumption in milligrams.
* **Computer Usage (min)**: Time spent on a computer in minutes.
* **Phone Usage (min)**: Time spent on a phone in minutes.
* **Total Sleep Duration (min)**: Total time spent asleep each night.
* **Sleep Onset Time**: Approximate time the subject went to sleep.
* **Custom Deep Sleep (hr)**: Estimated hours of deep sleep based on total sleep and influencing factors.
* **Sleep Efficiency (%)**: Percentage of time spent in bed that was actually spent sleeping.
* **Total Exercise (min)**: Sum of fitness and walking converted to an equivalent duration.
* **Total Screen Time (min)**: Combined phone and computer usage.
* **Protein Ratio**: Ratio of protein to total calorie intake.
* **Sugar-to-Carb Ratio**: Ratio of sugar to total carbohydrate intake.
* **Deep Sleep Ratio**: Ratio of deep sleep to total sleep.
* **Sleep Quality Score**: Final score (0–100) generated based on deep sleep, sleep timing, screen time, and caffeine impact.

1. **Data Collection**

The data used in this study was collected manually on a daily basis between March 15, 2025, and May 20, 2025. The dataset consists of 67 consecutive daily records. Each entry includes various lifestyle, nutrition, and digital behavior metrics related to sleep quality.

The data sources and methods include:

* Self-reported nutrition intake such as carbohydrates, sugar, protein, caffeine, and total calorie consumption, based on food labels and mobile nutrition tracking apps.
* Physical activity data including daily step counts and fitness exercise durations were collected using smartphone pedometer apps and personal fitness logs.
* Screen time (computer and phone usage) was tracked using built-in digital well-being tools available on smartphones and operating systems.
* Sleep data, such as total sleep duration and estimated deep sleep, was manually recorded each morning based on smartwatch outputs and subjective reporting. A calculated score, *Sleep Quality Score*, was derived by combining various influential factors like screen time, caffeine, and sleep onset time.

In total, 21 parameters were compiled for each day to enable both statistical and machine learning analysis. The final dataset was prepared and structured in Excel format and imported into Google Colab for analysis.

1. **Exploratory Data Analysis**

Before moving into statistical testing and machine learning modeling, a detailed exploratory data analysis (EDA) was conducted to understand the patterns and relationships hidden in the data. This phase involved the use of correlation matrices, scatter plots, boxplots, and time series visualizations to explore how lifestyle factors—such as exercise, caffeine intake, sugar consumption, screen time, and sleep schedule—affect sleep quality and deep sleep duration. Multiple variables were visually and statistically compared with key sleep indicators like **Sleep Efficiency (%)** and **Custom Deep Sleep (hr).** We tested different groupings (e.g., early vs late sleepers, high vs low caffeine consumers), examined linear relationships, and normalized variables to visualize trends over time. These observations informed the hypothesis testing and guided the selection of features for model training in the next phases.

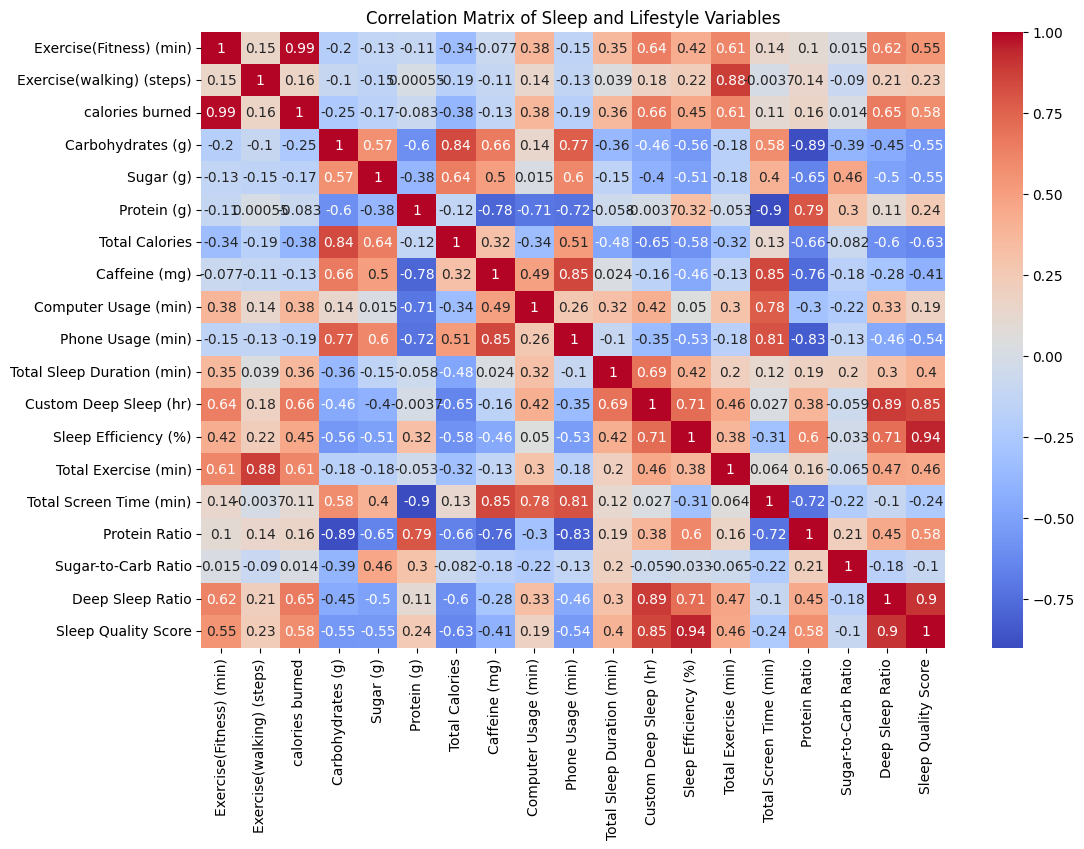
**Correlation Matrix of Sleep and Lifestyle Variables**

The correlation matrix provides an overview of linear relationships between key lifestyle and sleep-related variables.

Several noteworthy patterns emerge:

* Sleep Quality Score and Sleep Efficiency (%) show strong positive correlations with Total Sleep Duration, Custom Deep Sleep, and Deep Sleep Ratio.
* Caffeine (mg), Sugar (g), and Total Screen Time (min) are negatively correlated with both sleep efficiency and deep sleep metrics, confirming assumptions made during initial hypotheses.
* Exercise (Fitness) (min) and Total Exercise (min) are positively associated with deep sleep and sleep quality, reinforcing the beneficial role of physical activity.
* Some features, such as Carbohydrates (g) and Total Calories, display moderate positive associations with sugar and caffeine intake, reflecting their nutritional interdependence.

This matrix helped identify which variables would be most informative for further statistical testing and machine learning modeling.



**Relationship Between Exercise and Sleep Quality**

The scatter plots above demonstrate the effect of Exercise (Fitness) (min) on two key outcomes:

* Sleep Efficiency (%): A positive but moderately dispersed pattern indicates that more fitness activity may lead to higher sleep efficiency.
* Deep Sleep Duration (hr): A clearer positive trend is observed—individuals engaging in more fitness minutes generally achieve greater amounts of deep sleep.

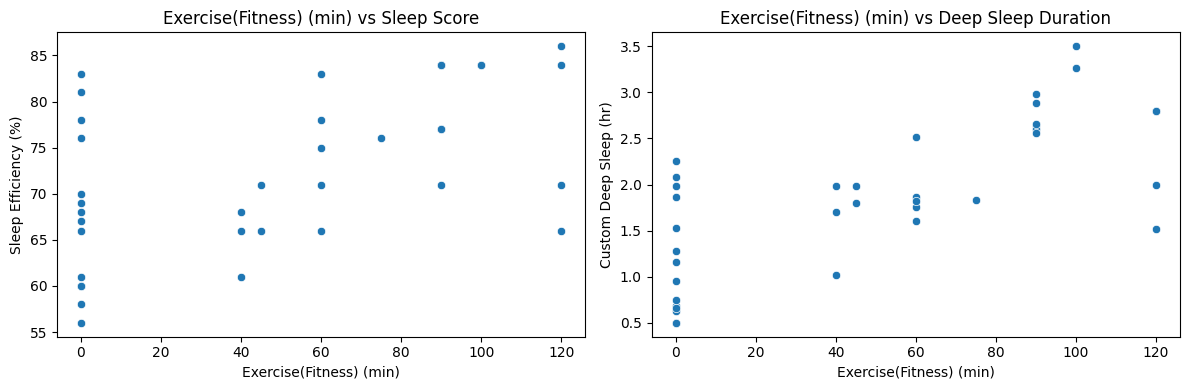
This supports the hypothesis that physical activity improves both the quality and depth of sleep.

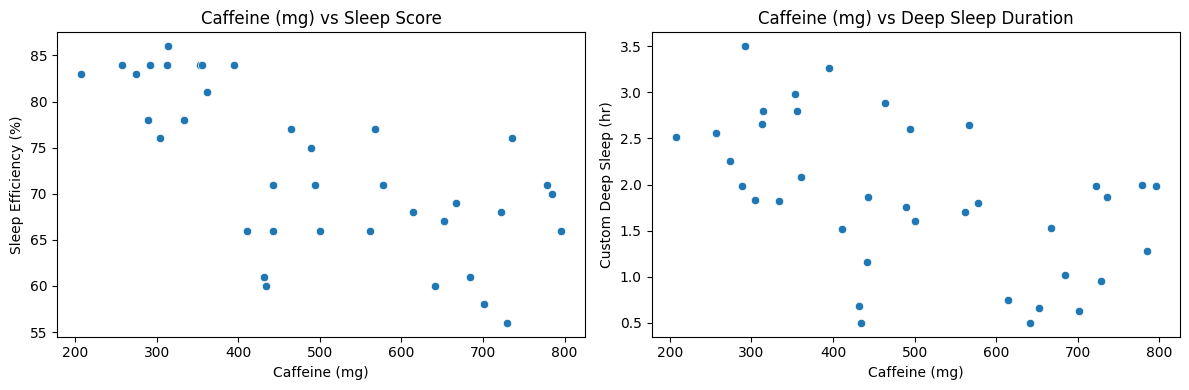
**Relationship Between Caffeine Intake and Sleep Quality**

The second set of plots visualizes the effect of Caffeine (mg) on:

* Sleep Efficiency (%): A noticeable downward trend is observed—higher caffeine levels are associated with lower sleep scores.
* Deep Sleep Duration (hr): The same negative relationship is visible here as well, especially above 500 mg, indicating that excessive caffeine consumption may impair deep sleep stages.

These results confirm the disruptive impact of caffeine on sleep and reinforce findings from correlation and hypothesis testing.





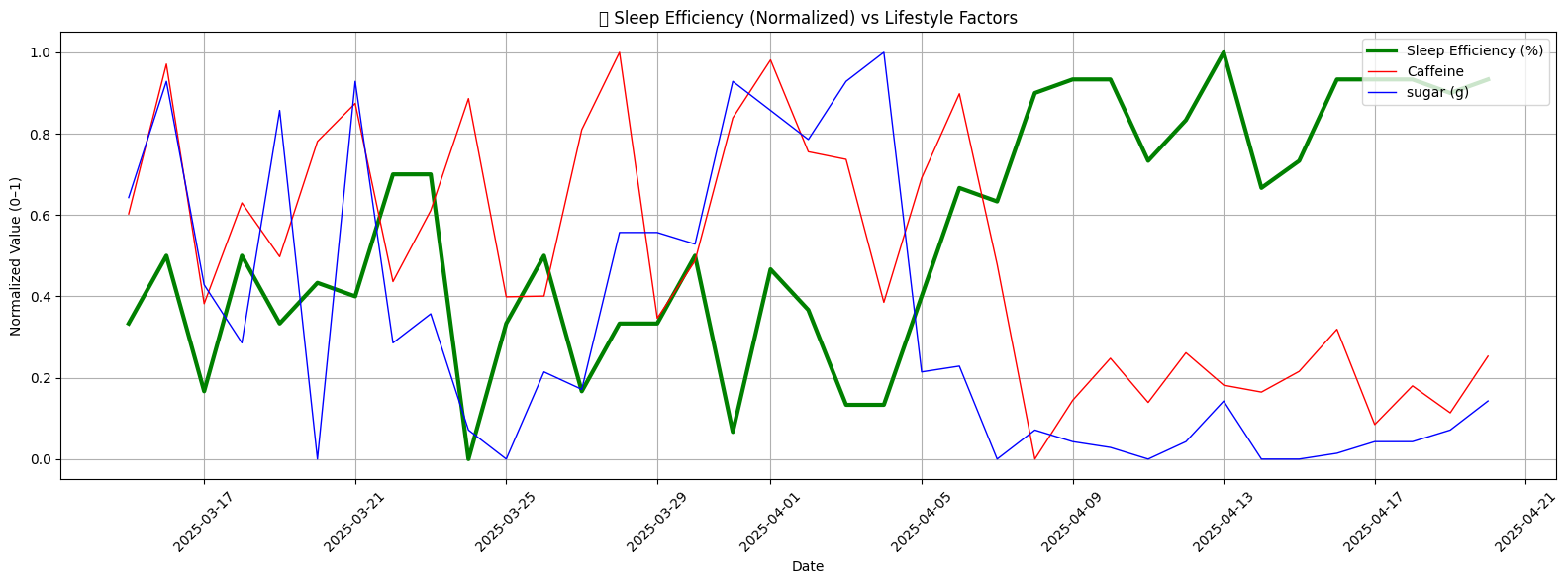
**Time Series: Sleep Efficiency vs Caffeine and Sugar Intake**

The line chart above presents a normalized comparison of Sleep Efficiency (%), Caffeine (mg), and Sugar (g) over time.

* The green line represents normalized sleep efficiency.
* The red line shows caffeine intake, while the blue line indicates sugar consumption.

A visible pattern can be observed where periods of high caffeine and sugar intake coincide with drops in sleep efficiency, particularly between late March and early April. Conversely, in the second half of the dataset—when caffeine and sugar levels decline—sleep efficiency improves steadily and remains consistently high.

This visual trend supports earlier correlation and hypothesis testing results that identified caffeine and sugar as negative influencers on sleep quality.



**H1 – Caffeine vs Sleep Duration**

This scatter plot examines the relationship between daily caffeine intake and total sleep duration (min). The negative slope indicates a slight downward trend; however, the Pearson correlation coefficient is r = -0.10 with a p-value = 0.54, meaning the result is not statistically significant.

Conclusion: While caffeine intake may appear to reduce sleep, this effect was not supported by significant evidence in this dataset.

**H2 – Exercise vs Deep Sleep**

The boxplot compares deep sleep duration between two groups: those who exercised (“Active”) and those who did not (“Inactive”).

The results show a clear difference, with the active group achieving significantly more deep sleep on average. The t-test result was t = 4.98, p < 0.001, indicating a highly significant effect of exercise on deep sleep.

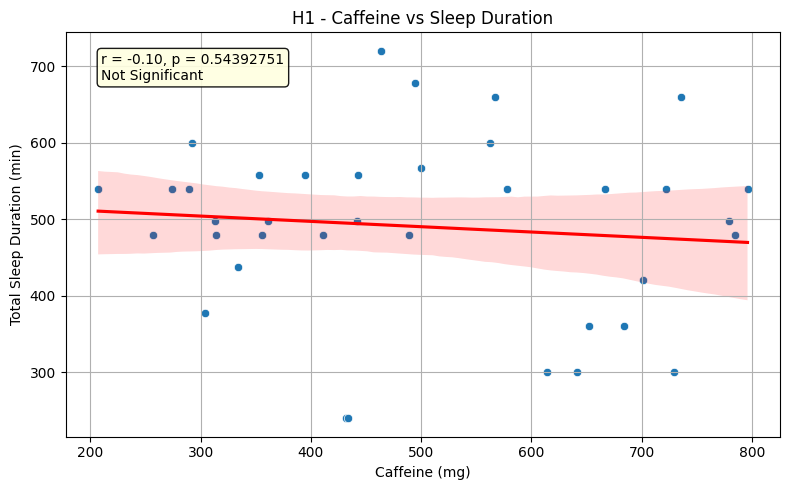
Conclusion: Regular physical activity has a positive and significant impact on deep sleep duration.

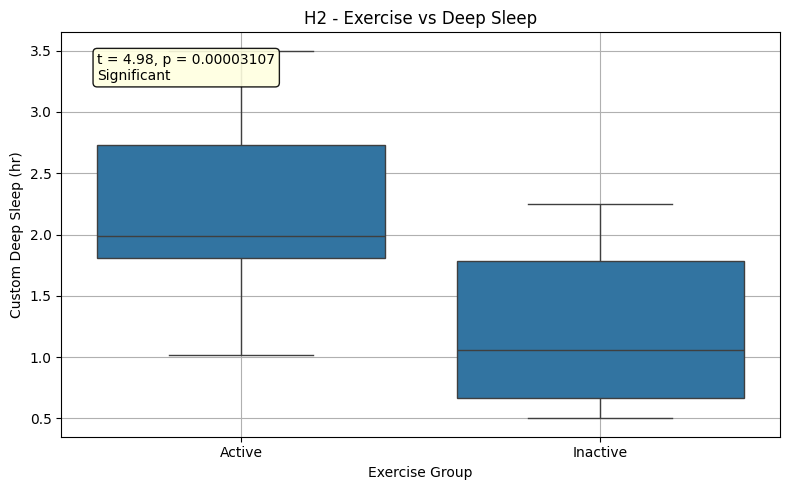
**H3 – Early vs Late Sleepers (Efficiency)**

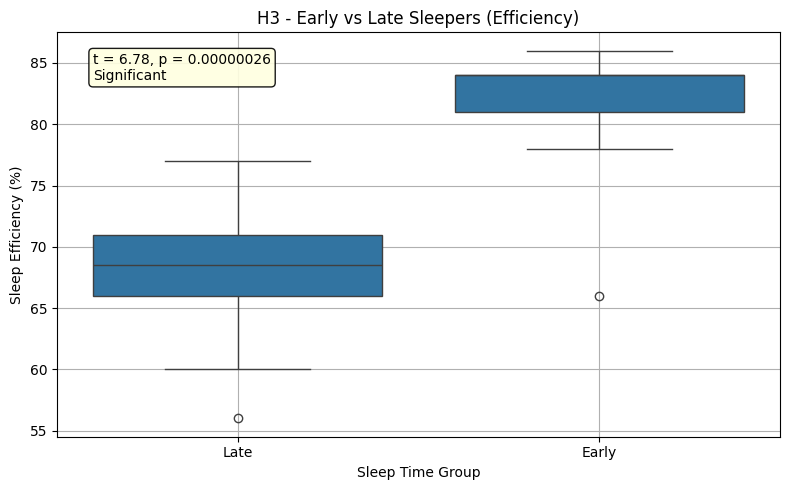
This boxplot contrasts sleep efficiency between those who went to bed early and those who slept late. The early sleepers had consistently higher efficiency scores.

The t-test yielded t = 6.78, p < 0.001, marking the difference as statistically significant.

Conclusion: Going to bed earlier is strongly associated with higher sleep efficiency.







1. **Machine Learning Modelling & Results**

Following the exploratory data analysis, we identified key variables that strongly influence sleep quality. However, raw features alone were not sufficient to capture the complex interactions between lifestyle behaviors and sleep outcomes. Therefore, a feature engineering phase was implemented to enhance the dataset with new, derived metrics that better represent these relationships. This step aimed to improve the model’s predictive accuracy and interpretability by creating meaningful variables—such as total activity levels, screen exposure, nutritional balance, and relative sleep quality measures. These engineered features were informed by statistical patterns observed during EDA and were selected to align with scientific literature on sleep and behavioral health. In the following section, we describe each newly created feature, its purpose, and how it contributes to the modeling process.

**Before Feature Engineering**

Initially, a Random Forest Regressor was trained using only the raw variables collected from daily behavior tracking such as fitness minutes, screen time, sugar intake, and sleep metrics.

The model performance metrics were:

* Mean Absolute Error (MAE): 2.80
* R² Score: 0.9371

These values suggest that the model was able to predict Sleep Quality Score with relatively high accuracy, even without engineered features.

metin, ekran görüntüsü, çizgi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

In the feature importance chart, the following variables emerged as the most predictive:

* Sleep Efficiency (%)
* Deep Sleep Ratio
* Custom Deep Sleep (hr)

Other variables such as screen time, sugar intake, and exercise minutes contributed to a lesser extent.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

The prediction plot shows a good alignment between predicted and actual sleep scores, indicating strong performance even before enhancing the data structure.

**After Feature Engineering**

After careful analysis, several new features were created to improve model performance and capture more nuanced patterns:

* Screen Load Ratio: screen time adjusted to sleep duration
* Caffeine-to-Exercise Ratio: caffeine levels relative to physical activity
* Sugar-to-Carb Ratio: proportion of sugar in total carbohydrate intake
* Sleep Load Score: composite score aggregating negative and positive sleep influencers

The same Random Forest model was retrained using these new features alongside the original ones.

**Updated model results:**

* Mean Absolute Error (MAE): 2.73
* R² Score: 0.9347

This indicates a slight improvement in residual error and a maintained high level of accuracy.

metin, ekran görüntüsü, çizgi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Post-engineering, Sleep Efficiency (%), Deep Sleep Ratio, and the newly created Screen Load Index became key drivers in predicting Sleep Quality Score.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

The new prediction plot also shows strong clustering along the identity line, confirming continued predictive alignment.

ekran görüntüsü, metin, ekran, görüntüleme, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Cross-validation confirmed the model’s robustness, with R² values ranging from 0.82 to 0.94 across different folds.

1. **Conclusion**

This project aimed to uncover the relationship between daily lifestyle behaviors and sleep quality through personal data collection, statistical analysis, and machine learning modeling. Throughout the analysis, we observed that variables such as **sleep efficiency, deep sleep ratio, caffeine intake,** and **screen time** had strong correlations with overall sleep performance. Hypothesis testing and visual trends supported the idea that high caffeine intake, late sleep onset, and excessive screen exposure negatively impact sleep, while regular physical activity and earlier bedtimes contribute positively. After developing new composite features (e.g**., Screen Load Ratio, Caffeine-to-Exercise Ratio, Sleep Load Score),** our **Random Forest model** achieved robust predictive performance (R² ≈ 0.93), validating the power of engineered indicators over raw data alone. This process not only provided a deeper understanding of personal sleep dynamics but also demonstrated the importance of structured data and thoughtful feature engineering in predictive analytics. The insights gained can now inform more data-driven lifestyle decisions aimed at improving rest, focus, and overall well-being.

1. Appendix

<https://openai.com/index/gpt-4/>

<https://play.google.com/store/apps/details?id=com.sleepmonitor.aio&hl=en>