Digital Health Report
Sleep Analysis

I. Introduction

Research has shown the complex relationship between sleep patterns and various lifestyle factors in recent years. This report dives into the sleep-related data gathered by the Sleep Cycle app, with an emphasis on factors impacting sleep quality. To provide context to the exploration, key findings are drawn from studies that investigated the effects of caffeine consumption, exercise, sleep disturbance, and stress levels on sleep outcomes.

Watson et al. (2016) conducted a cross-sectional study in Australian adults to investigate the relationship between caffeine use and sleep quality. This study provides a basic understanding of how coffee, a common component of daily life, affects sleep measurements. Their results demonstrated a strong relationship between increased coffee consumption and decreased time in bed. Moreover, individuals who reported poor sleep quality consumed much more caffeine than those who reported good sleep quality.

Kelley and Kelley (2017) add to our understanding of the positive effects of exercise on sleep outcomes in adults. They showed statistically significant enhancements in multiple sleep indices, such as the apnea-hypopnea index, overall sleep quality, subjective sleep, and sleep latency. This was determined through a meta-analysis study. The findings of this research lay the foundation for future research on the benefits of physical activity in enhancing sleep quality. This aligns with the project's goal of evaluating the impact of exercise on sleep quality.

Herawati and Gayatri (2019) investigated the relationship between sleep quality and stress levels in college students. Their cross-sectional study found a strong association, indicating that students with poor sleep quality were 4.7 times more likely to be stressed. This study serves as an important reminder of the connection between sleep quality and psychological well-being. This is also in line with the purpose of this project, which is to investigate the relationship between sleep quality and reported stress levels.

Bin et al. (2016) investigated the relationship between sleep disruption and the development of health risks, particularly hypertension and dyslipidemia. Their large longitudinal study of over 45,600 adults found that sleep disturbance predicted an increased risk of hypertension and dyslipidemia. This study also emphasizes the idea that sleep quality is more relevant than sleep quantity in predicting health risks. These findings prompt further investigation into the health implications of sleep quality in our dataset.

This report builds on existing literature, exploring the impacts of caffeine, exercise, sleep disruption, and stress on sleep quality. Hence, the primary research question is: **How do lifestyle factors, including caffeine consumption, stress levels, and exercise habits, influence sleep quality, moving beyond the conventional focus on sleep duration?** Moreover, the report aims to gain insights from clustering individuals based on their sleep patterns and determine the accuracy of predicting sleep quality from a set of features.

The research uses a systematic approach, starting with data cleaning to ensure the reliability of the dataset. Initial observations on the subject level involve analyzing average sleep quality trends over time. Techniques such as correlation matrices are then used to investigate correlations among sleep quality, time in bed, heart rate, and activity. Exploratory data analysis then delves into specific lifestyle factors such as caffeine consumption, exercise habits, and stress levels. Clustering analysis, specifically K-Means clustering, is applied to identify distinct sleep patterns based on relevant variables. Lastly, machine learning models, such as linear regression, logistic regression, and random forest classifiers, are implemented to predict sleep quality.

Observations on individuals revealed trends and correlations, adding to the overall understanding of sleep patterns. Exploratory data analysis hinted at connections between caffeine, exercise, stress, and sleep quality. Tea generally yields higher sleep quality than coffee, exercise is linked to improved sleep, and unexpectedly, coffee consumption and stress are associated with higher sleep quality. The study discovers two sleep patterns using clustering analysis, showing how lifestyle factors influence sleep quality. Lastly, the machine learning models did not offer accurate predictions, likely due to variations in the data.

II. Problem Formulation

The research problem in this investigation is centered around understanding the intricate relationship between lifestyle factors and sleep quality. Unlike conventional studies that primarily concentrate on sleep duration, this project delves into the less-explored realm of quality of sleep. Therefore, the formulation addresses a gap in current research. This shift in focus aligns with research findings, such as those presented by Bin et al. (2016), highlighting the critical role of sleep quality in predicting health risks.

The main research question focuses on the influence of lifestyle factors, such as caffeine consumption, stress levels, and exercise habits, on sleep quality. Understanding how our lifestyle choices affect our sleep quality is important if we want to improve our sleep. The goal is to go beyond the usual emphasis on sleep duration. The clustering method helps categorize different sleep patterns, providing specific insights for better sleep. Moreover, the study aims to determine if accurate predictions about sleep quality can be made based on a set of characteristics. Additionally, the use of machine learning to predict sleep quality can contribute to the development of user-friendly tools for personalized sleep management.

The research question delves into the relationship between lifestyle and sleep quality by closely examining both individuals and the community. This not only improves our understanding of this relationship but also meets the requirements for a detailed study.

III. Dataset Description

The dataset consists of sleep-related data collected over the period of 2014-2018, originating from the Sleep Cycle app developed by Northcube for iOS. The dataset contains 887 observations, each capturing essential metrics related to an individual's sleep patterns. These observations are represented by timestamped entries, denoting the start and end of each sleep session. Presented below is a detailed breakdown of the columns within the dataset:

- Start **[object]**: The timestamp indicating the start of each sleep session, in the format YYYY-MM-DD HH:MM:SS.
- End **[object]**: The timestamp indicating the end of each sleep session, also in the format YYYY-MM-DD HH:MM:SS.
- Sleep Quality **[object]**: A percentage value representing the quality of sleep experienced. This value can range from 0% (poor quality) to 100% (excellent quality).
- Time in Bed **[object]**: The duration of time the user spent in bed during the sleep session, represented in hours and minutes.
- Wake up **[object]**: The user's emotional state or mood upon waking up, which is conveyed through emoticons as follows:
    - :) (Smiling Face): indicates a positive or happy mood upon waking up.
    - :| (Neutral Face): indicates a neutral or average emotional state upon waking.
    - :((Frowning Face): indicates a negative or unhappy mood upon waking up.

- Sleep Notes **[object]**: Additional notes that include information about stress, caffeine consumption, work out, and late-night eating.
- Heart Rate **[float64]**: The user's heart rate during the sleep session, in beats per minute (BPM)
- Activity (Steps) **[int64]**: The number of steps or activity recorded during the day.

This dataset offers valuable insights into an individual's sleep patterns and how various factors may affect the quality and duration of their sleep. Each entry in the dataset represents a distinct sleep session. The dataset features various types of data, including quantitative data such as 'Sleep Quality,' 'Time in Bed,' 'Heart Rate,' and 'Activity (Steps).' Additionally, categorical information is conveyed through the 'Wake up' column and the 'Sleep Notes' column.

Upon closer examination of missing values, it is observed that 'Wake up,' 'Sleep Notes,' and 'Heart Rate' contain some missing data. Specifically, 'Wake up' has 641 missing entries, 'Sleep Notes' lacks information in 235 instances, and 'Heart Rate' has 725 missing values. These missing values were addressed during data preprocessing. However, the frequency of data is not explicitly mentioned and may require further exploration.

The analysis focuses on several key features, including Sleep Quality, Time in Bed, Sleep Notes, Heart Rate, and Activity (Steps). These features were selected based on their relevance to sleep patterns and associated lifestyle factors. Specifically, 'Time in Bed,' 'Sleep Quality,' and 'Sleep Notes' are important for understanding sleep patterns and lifestyle factors.

While preparing the sleep dataset for analysis, several data preprocessing steps were implemented. First, the timestamps indicating the start and end of each sleep session were converted to a standard datetime format. The 'Time in Bed' column, initially represented as hours and minutes, was transformed by subtracting the 'Start' time from the 'End' time for each sleep session. The time duration was converted to seconds, and the data type was changed to 'timedelta64[s]'. This conversion helps in accurately analyzing sleep patterns over time.

Next, missing values were replaced with data to maintain the dataset's completeness. For the 'Wake up' feature, which reflects the user's mood upon waking up, missing entries were filled with the mode. The mode represents the most frequently occurring emotional state. Similarly, for 'Heart rate,' missing values were filled with the average heart rate, ensuring low bias in the data.

To maintain consistency, data cleaning was implemented. In the 'Sleep quality', the percentage sign was removed and converted into a simple number format. Additionally, 'Time in bed' was converted to seconds for uniformity, making it easier to compare across different timeframes.

New binary categorical features were created based on the 'Sleep Notes' column. This methodology was inspired by Alistair Acheson's insightful code shared on the Medium platform. Features such as ‘Drank tea,’ ‘Drank coffee,’ ‘Worked out,’ ‘Ate late,’ and ‘Stressful day’ help us understand how specific events might relate to sleep quality. This categorical information adds depth to the dataset, allowing for a deeper correlation analysis

Finally, irrelevant columns like 'Start,' 'End,' 'Wake Up' were dropped to streamline the dataset for easier analysis. These preprocessing steps ensure that the dataset is well-prepared for exploring the connections between sleep patterns, lifestyle choices, and overall well-being.

IV. Methods

In the analysis of the dataset, a range of methods were used to study sleep patterns and their connections to various lifestyle factors. Beginning with descriptive analysis, monthly averages and correlation matrices were used. This is to understand how sleep quality changes over time and to explore relationships between key variables like sleep quality, time in bed, heart rate, and activity.

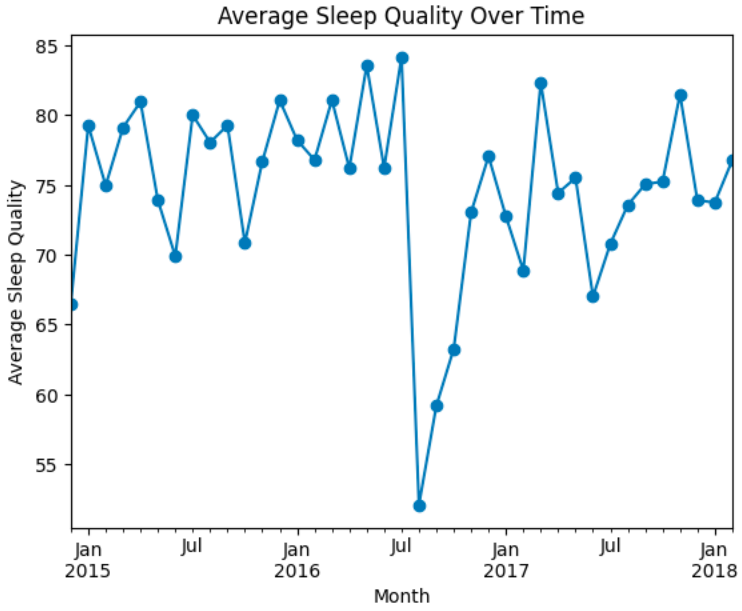

Figure 1: Line plot displaying the average sleep quality over consecutive months

The first method involves a detailed examination of sleep quality trends over time. This is achieved by creating a new feature, 'Month,' derived from the 'Start' timestamp, and then calculating the average sleep quality for each month. This method allows us to observe trends and patterns in sleep quality over monthly intervals. Understanding how sleep quality varies over time is crucial for identifying potential long-term influences or seasonal variations.

In Figure 1, the plot shows a periodic trend with gradual improvements and declines in sleep quality across months. However, during the period from July to November 2016, there is a steep decrease in the average sleep quality. Throughout this timeframe, the subject consistently experienced lower sleep quality, probably due to external factors such as lifestyle changes. During these months, there might be seasonal variations affecting sleep. For

instance, changes in temperature, daylight duration, travel, and time zone differences could disrupt their circadian rhythm and result in changes in sleep patterns.

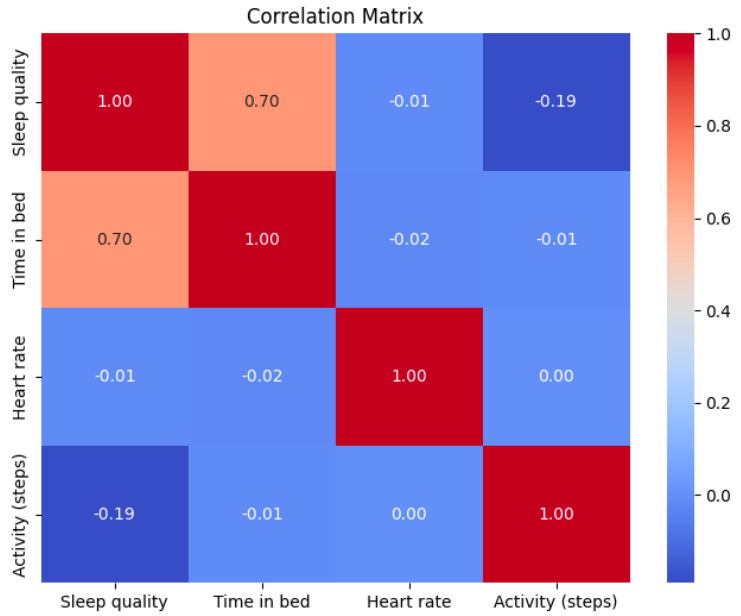

Figure 2: Correlation heatmap between 'Sleep Quality', 'Time in bed,' 'Heart rate,' and 'Activity (steps)

Next, a Correlation Matrix was created to provide a quantitative understanding of how these variables relate to each other. Understanding correlations helps identify potential relationships between sleep quality and other factors. The correlation matrix indicates a high positive correlation (0.7) between 'Sleep quality' and 'Time in bed'. The high positive correlation signifies that as the duration of sleep increases, sleep quality improves. However, the correlation with other variables, such as 'Heart rate' and 'Time in bed' are low and negative. This suggests weak linear associations between sleep quality and these variables.

The results from the line plot serve as a foundation for deeper investigations. If certain months consistently have lower sleep quality, it creates questions about potential external factors influencing sleep. Additionally, the strong correlation opens opportunities for a detailed exploration of the impact of lifestyle choices on sleep quality.

Therefore, to investigate lifestyle influences, binary categorical features such as 'Drank tea' or 'Worked out,' were created. Bar charts can visualize the average sleep quality in relation to distinct lifestyle factors. This method allows a direct comparison of average sleep quality tied to various lifestyle factors, offering a clear and simple depiction of the relationships.

However, the weak correlations highlight the need for a more nuanced exploration of the relationships between other variables. Further methods, such as clustering or machine learning, can be utilized to uncover non-linear patterns that might not be apparent in the correlation matrix.

Hence, K-Means clustering is applied to identify distinct groups with similar sleep patterns, presented by scatter plots. Machine learning methods, including linear regression, logistic regression, and a random forest classifier, were used to predict sleep quality based on various features. These methods were used to achieve a thorough understanding of the dataset. Additionally, it provides a quantitative understanding of the relationships.

V. Results

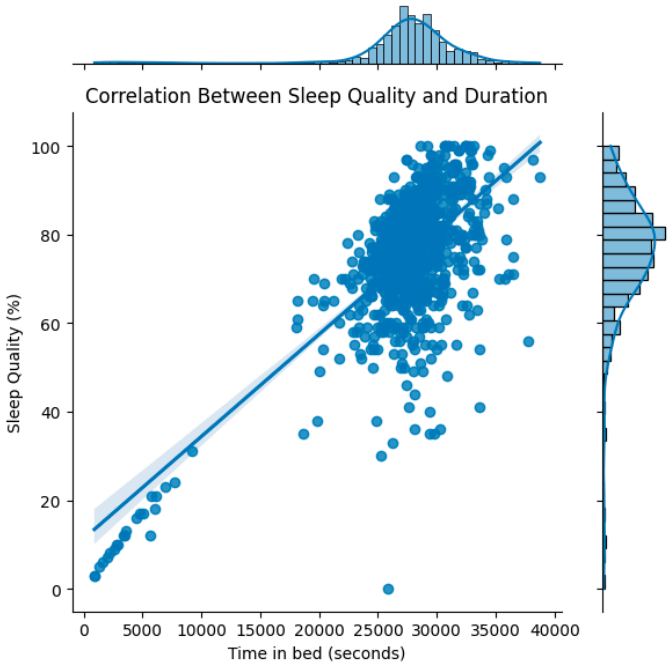

Figure 3: Joint plot with regression line between 'Sleep Quality' and 'Time in bed'

The joint plot with a regression line illustrates a linear proportional increase, indicating a positive correlation between two variables. In this context, it may be representing the linear relationship between the time spent in bed and sleep quality. The positive slope of the regression line indicates that as the duration of sleep increases, the sleep quality rises linearly.

For lifestyle choices like coffee consumption, tea consumption, working out, and experiencing stress, individual bar plots were generated to explore their impact on sleep quality.

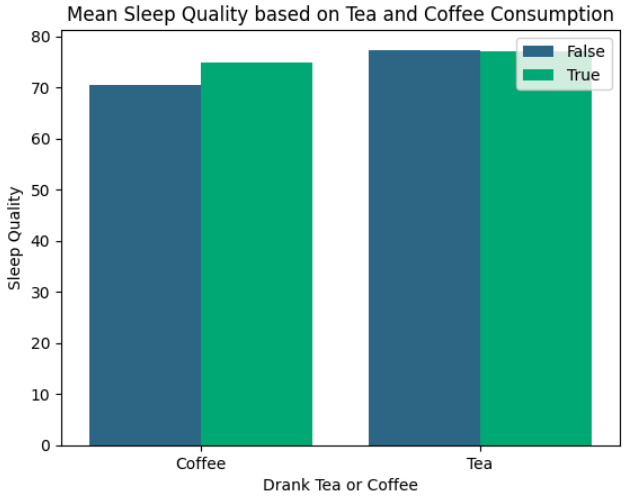

Figure 4: Bar plot of average sleep quality and caffeine intake

In contrast to Watson et al.'s (2016) conclusions, Figure 4 reveals an unexpected finding. On average, sleep quality appears to be higher after the consumption of coffee (74.85) compared to instances without coffee intake (70.52). Furthermore, the data suggests that sleep quality is even higher when tea is consumed (77.37). This supports popular beliefs regarding tea's calming impact on sleep. However, it is important to note that the results may be influenced by individual variations and requires further investigation.

Figure 5: Bar plot of average sleep quality and exercise

The data presented in Figure 5 indicates that sleep quality is generally higher following a workout (75.62) compared to periods without exercise (74.39). This result alligns with the findings of Kelley and Kelley's (2017) study, which links regular exercise to improved sleep quality.

Figure 6: Bar plot of sleep quality and stress

In contrast to the findings by Herawati and Gayatri (2019), the data in Figure 6 suggests that sleep quality tends to be higher when individuals experience stress (77.88) compared to periods without stress (74.69). This unexpected outcome could be due to the subjective nature of stress perception or may be influenced by confounding variables. The disparity in findings highlights the complexities of the interaction between stress and sleep quality.

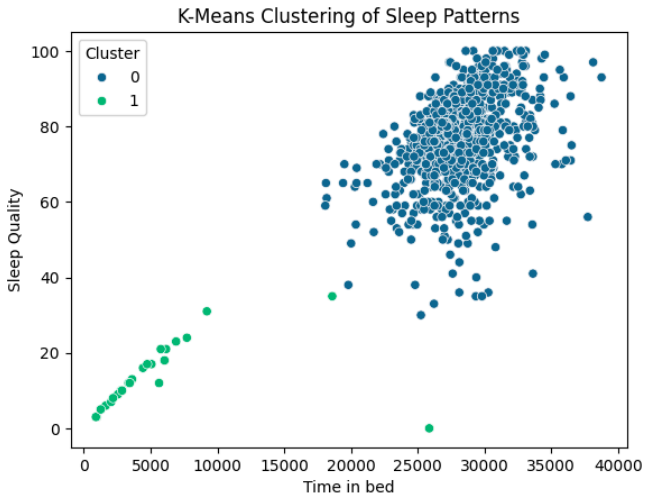

Figure 7: K means clustering

In the scatterplot generated by K-Means clustering, data points are grouped based on 'Time in bed' and 'Sleep quality'. The clusters are well-separated, indicating distinct groups with similar sleep patterns. Cluster 1 (green) could be linked with shorter sleep duration and lower sleep quality, while Cluster 0 (blue) indicates longer sleep duration and higher sleep quality. Moreover, the clusters are not evenly distributed in terms of size. Cluster 0 appears denser with data points closer together, while Cluster 1 is less compact. Majority of data points align with characteristics associated with Cluster 0. This emphasizes a trend of longer sleep duration and higher sleep quality.

Lastly, machine learning models were used to predict sleep quality based on various features. The accuracy results show that linear regression performed the best, with a test accuracy of approximately 41.94%. However, both logistic regression and the random forest classifier produced lower accuracy values of around 5.06% and 3.93% respectively.

VI. Conclusion & Discussion

In conclusion, this study of the sleep dataset has provided important insights into the factors affecting sleep patterns and quality. The descriptive methods allowed us to identify temporal trends and relationships between variables. Clustering techniques revealed potential patterns in sleep behavior, while the machine learning models provided predictive insights. Additionally, the correlation matrix showed a strong positive connection between sleep quality and the time spent in bed, emphasizing the importance of sleep duration in determining overall sleep quality.

However, the low negative correlations observed between certain variables, such as heart rate or activity, may be due to limited data. Despite the observed trends, there was a drawback in the dataset. Among the 887 entries, the heart rate data was missing for a significant 725 instances, leaving a considerable gap in the dataset. To handle this issue, a practical approach was to replace the missing heart rate values with the average heart rate. While this allowed the inclusion of heart rate in the analysis, it introduced biases and uncertainties. The lack of heart rate details decreased the ability to thoroughly investigate its impact on sleep quality.

For future studies, it would be beneficial to explore more advanced ways of filling in missing values.

The results provide intriguing insights into how lifestyle factors affect sleep quality, while challenging some conventional expectations. Generally, tea appears to contribute to higher sleep quality when compared to coffee. The study also supports the established link between engaging in a workout and improved sleep quality. However, the data reveals an unexpected trend where sleep quality is higher with coffee consumption (74.85) than without (70.52). Additionally, the study contradicts previous research by suggesting higher sleep quality during stressful periods (77.88) rather than stress-free periods (74.69).

It is important to note that the differences in sleep quality are relatively modest. The divergence from expectations based on existing research highlights the need for further investigation into confounding variables. Factors such as variations in caffeine content, individual differences in caffeine metabolism, and the complexity of lifestyle factors could contribute to these differences. Future research should explore these factors to uncover the complex relationships between lifestyle choices and sleep quality.

Furthermore, the interpretability of clusters is subjective, and the effectiveness of clustering depends on the chosen features. While the clusters in this study provided insights into sleep patterns, the uneven cluster sizes suggest potential imbalances in data distribution. The clustering algorithm's sensitivity to outliers may also affect the accuracy of group assignments.

Finally, the machine learning models used to predict sleep quality showed lower accuracies than expected. These reduced accuracies could be due to the complexities in the relationships between the features and sleep quality. Additionally, the models may have missed non-linear patterns in the data, resulting in poor prediction performance. To improve future models, advanced algorithms can be used to understand complex relationships in the data.

For future analyses, incorporating more diverse datasets, including longitudinal studies and objective sleep measurements, could enhance findings. Additionally, investigating the influence of external factors, such as environmental conditions and individual health profiles, could contribute to a more detailed exploration of sleep.

VII. References

Acheson, A. (2021). *Sleep iOS App Data — Exploratory Analysis with Python*. [online] Medium. Available at: https://medium.com/@acheson.alistair/sleep-ios-app-data-exploratory-analysis-with-python-d2cd9eef5a8f [Accessed 11 Dec. 2023].

Bin, Y.S. (2016). Is Sleep Quality More Important than Sleep Duration for Public Health? *Sleep*, 39(9), pp.1629–1630. doi:https://doi.org/10.5665/sleep.6078.

Herawati, K. and Gayatri, D. (2019). The correlation between sleep quality and levels of stress among students in Universitas Indonesia. *Enfermería Clínica*, [online] 29(2), pp.357–361. doi:https://doi.org/10.1016/j.enfcli.2019.04.044

Kelley, G.A. and Kelley, K.S. (2017). Exercise and sleep: a systematic review of previous meta-analyses. *Journal of Evidence-Based Medicine*, 10(1), pp.26–36. doi:https://doi.org/10.1111/jebm.12236.

Watson, E., Coates, A., Kohler, M. and Banks, S. (2016). Caffeine Consumption and Sleep Quality in Australian Adults. *Nutrients*, [online] 8(8), p.479. doi:https://doi.org/10.3390/nu8080479.

www.kaggle.com. (n.d.). *Sleep Data*. [online] Available at: https://www.kaggle.com/datasets/danagerous/sleep-data.