Table of Contents

[Answer to Q1: 2](#_Toc163849142)

[Answer to Q2: 7](#_Toc163849143)

[Answer to Q3: 8](#_Toc163849144)

[Answer to Q4: 10](#_Toc163849145)

[Answer to Q5: 14](#_Toc163849146)

[Answer to Q6: 17](#_Toc163849147)

[Answer to Q7: 18](#_Toc163849148)

[References 19](#_Toc163849149)

# Answer to Q1:

**Data Exploration**

This dataset (“sleep”) contains **16 variables** and **452 rows**, with **no missing values**. However, after a quick scan of the dataset, there were two notable observations that stood out:

1. The values in **‘**Bedtime’ and ‘Wakeup time’ are flawed and do not tally. Zooming in on individual 3 (Figure 1), both dates in ‘Bedtime’ and ‘Wakeup time’ are recorded as 5/25/2021. However, with a bedtime of 21:30 and a sleep duration of 8 hours, the wakeup time should logically be on 5/26/2021 05:30. This inconsistency happens throughout the dataset, hence rendering the dates in these columns purposeless.

A screenshot of a computer

Description automatically generated

Figure 1. First 10 records of sleep.csv

1. 2 individualsunder the age of 21 had a smoking status of Yes (Figure 2), which is illegal in both Singapore and the United States. However, it is not entirely known which country this dataset was surveyed from, hence we will assume that the legal age for tobacco products is 18 and keep these rows in our dataset.



Figure 2. Records of two 19-year-old individuals who smoke

The following data cleaning steps were done in R (Figure 3):

1. Used ‘janitor’ library and clean\_names() function to standardise the naming convention of the variables *(i.e. ‘Smoking status’ to ‘smoking\_status’)*
2. Dropped ‘id’ variable
3. Changed ‘gender’ and ‘smoking\_status’ to factors
4. Dropped the dates from ‘bedtime’ and ‘wakeup\_time’ variables
5. Changed ‘bedtime’ and ‘wakeup\_time’ to factors
6. Added an ‘age\_group’ variable that categorises the different ages into age groups, as follows:

|  |  |
| --- | --- |
| **Ages** | **Age Group** |
| 9 – 12 | Children |
| 13 – 17 | Adolescents |
| 18 – 24 | Young Adults |
| 25 – 39 | Adults |
| 40 – 59 | Middle-aged Adults |
| 60 – 69 | Older Adults |

A computer code with blue text

Description automatically generated

Figure 3. Data cleaning procedures in R

After cleaning the data, the variables and datatypes we will be working with are shown in the table and Figure 4 below.

|  |  |  |
| --- | --- | --- |
| **No.** | **Variable** | **Datatype** |
| 1. | age | Integer |
| 2. | gender | Factor |
| 3. | bedtime | Factor |
| 4. | wakeup\_time | Factor |
| 5. | sleep\_duration | Numeric |
| 6. | sleep\_efficiency | Numeric |
| 7. | rem\_sleep\_percentage | Integer |
| 8. | deep\_sleep\_percentage | Integer |
| 9. | light\_sleep\_percentage | Integer |
| 10. | awakenings | Integer |
| 11. | caffeine\_consumption | Integer |
| 12. | alcohol\_consumption | Integer |
| 13. | smoking\_status | Factor |
| 14. | exercise\_frequency | Integer |
| 15. | daily\_steps | Integer |
| 16. | age\_group | Factor |

A close-up of a table

Description automatically generated

Figure 4. Summary statistics of processed dataset

**Key Finding 1 (Sleep Efficiency VS Gender + Age Group)**

The boxplot illustrates the sleep efficiency across different genders and age groups.

A graph of different colored rectangles

Description automatically generated

Figure 5. Boxplot of sleep efficiency across gender and age groups

According to the visualisation, these are the observations:

1. **Sleep efficiency** generally **increases with age** for **both genders** before it dips once it reaches the 60 – 69 age range.
2. **Females** tend to have **slightly higher median sleep efficiency than males** across **most age groups** (young adults, adults, and middle-aged adults), apart from **older adults.**
3. **Middle-aged females** have the **highest median sleep efficiency**, which could be due to factors such as lifestyle changes or hormonal changes.

It is important to note that these are just observational findings from a chart, and more research is needed to confirm these patterns and understand the underlying causes.

**Key Finding 2: Correlation of Continuous Variables**

The correlation plot illustrates the correlation of all continuous variables in the dataset.

A screen shot of a graph

Description automatically generated

Figure 6. Correlation plot of continuous variables

According to the visualisation, these are the observations:

1. **Sleep Efficiency and Deep Sleep Percentage (0.79):**

Individuals with a higher sleep efficiency tend to spend a greater proportion of their total sleep time in deep sleep stages. Conversely, lower sleep efficiency may be associated with a lower percentage spent in deep sleep.

1. **Sleep Efficiency and Light Sleep Percentage (-0.82):**

Individuals with a higher sleep efficiency tend to spend less time in light sleep stages. Conversely, lower sleep efficiency may be associated with a higher percentage spent in light sleep.

1. **Sleep Efficiency and Awakenings (-0.54)**:

Individuals with higher sleep efficiency tend to experience fewer awakenings during the night. Conversely, lower sleep efficiency may be associated with a higher frequency of awakenings, potentially disrupting the overall sleep continuity.

**Key Finding 3: Optimal Bedtime at 10.30PM**

This boxplot illustrates the sleep efficiency across different bedtimes from 21:00 to 02:30 and the mean and median sleep efficiency has been extracted and put in a table below for easier interpretation.

A chart with different colored rectangles

Description automatically generated

Figure 7. Boxplot of sleep efficiency across different bedtimes

|  |  |  |
| --- | --- | --- |
| **Bedtime** | **Mean Sleep Efficiency** | **Median Sleep Efficiency** |
| 21:00 | 0.823 | 0.835 |
| 21:30 | 0.792 | 0.820 |
| 22:00 | 0.810 | 0.865 |
| 22:30 | 0.849 | 0.870 |
| 23:00 | 0.785 | 0.850 |
| 00:00 | 0.782 | 0.820 |
| 00:30 | 0.800 | 0.860 |
| 01:00 | 0.723 | 0.740 |
| 01:30 | 0.799 | 0.840 |
| 02:00 | 0.753 | 0.720 |
| 02:30 | 0.765 | 0.795 |

According to the visualisation and table, these are the observations:

1. **Bedtimes around 22:30** appear to be associated with the **highest average (0.849)** and **median sleep efficiency (0.870)**. This indicates that individuals who go to bed around 22:30 tend to have better sleep quality compared to those with other bedtimes.
2. Average sleep efficiencies are **lower** for individuals **who sleep later than 01:00**, ranging from **0.723 to 0.799**.

However, it is important to note that the individuals’ personal lifestyles do have an impact on contributing to sleep efficiency, therefore, 22:30 may be the optimal bedtime according to this visualisation, it does not tell the full story and require a deeper analysis.

# Answer to Q2:

Research Question 1

Can we predict sleep efficiency based on an individual’s lifestyle factors and habits, while controlling for demographic variables like age and gender?

Research Question 2

Can we predict sleep efficiency solely based on physiological sleep metrics, such as REM sleep percentage, deep sleep percentage, light sleep percentage and awakenings without considering lifestyle factors and demographic variables?

# Answer to Q3:

Research Question 1

An individual’s lifestyle factors and habits comprise the following variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Bedtime | Wakeup Time | Sleep Duration | Caffeine Consumption |
| Alcohol Consumption | Smoking Status | Exercise Frequency | Daily Steps |

On top of these 8 variables, the demographic variables ‘age’ and ‘gender’ are included in the consideration.

A linear regression model with the 10 mentioned variables, and with ‘sleep\_efficiency’ as the target variable was run. Subsequently, to remove the statistically insignificant variables, we ran stepwise regression (Figure 8), which is a method that “iteratively examines the statistical significance of each independent variable in a linear regression model” (Hayes, 2022).

A screenshot of a computer

Description automatically generated

Figure 8. Stepwise model results

The selected target variable and input variables were chosen because they were identified as the most statistically significant variables related to lifestyle factors and habits through a stepwise regression analysis, ensuring the inclusion of the most influential predictors for analysing sleep efficiency. These variables are shown in the table below.

|  |  |
| --- | --- |
| Can we predict sleep efficiency based on an individual’s lifestyle factors and habits, while controlling for demographic variables like age and gender? | |
| **Target Variable** | Sleep Efficiency |
| **Input Variables** | Bedtime |
| Alcohol Consumption |
| Smoking Status |
| Exercise Frequency |
| Daily Steps |
| Age |

Research Question 2

Considering that there is a strong correlation between sleep efficiency and deep sleep percentage from data exploration, it is important to find out if it is possible to predict sleep efficiency solely based on physiological sleep metrics instead of external factors. These physiological sleep metrics include the following variables:

|  |  |  |  |
| --- | --- | --- | --- |
| REM Sleep Percentage | Deep Sleep Percentage | Light Sleep Percentage | Awakenings |

Similar to Research Question 1, I ran a linear regression model (Figure 9) and realised Light Sleep Percentage was not defined because of singularities, whereas the other variables were of high statistical significance.

A screenshot of a computer

Description automatically generated

Figure 9. Linear regression model for Q2

To verify this, I ran it through a stepwise regression, which leaves us with the target variable and input variables as seen from the table below.

|  |  |
| --- | --- |
| Can we predict sleep efficiency solely based on physiological sleep metrics, such as REM sleep percentage, deep sleep percentage, light sleep percentage and awakenings without considering lifestyle factors and demographic variables? | |
| **Target Variable** | Sleep Efficiency |
| **Input Variables** | REM Sleep Percentage |
| Deep Sleep Percentage |
| Awakenings |

# Answer to Q4:

To analyse the predictive performance of different models consistently across the questions, I utilised the same methodology for both research questions. I filtered the statistically significant variables into a new data frame before doing a 70-30 train-test split (Figure 10). This involved splitting the data into training and test sets. The training set was used to build the models, while the testing set was used to evaluate the performance on unseen data.

A computer screen shot of text

Description automatically generated

Figure 10. Train-test split

Following this split, I built a Random Forest model first for each question, then followed by building a Multivariate Adaptive Regression Splines (MARS) model. For the MARS models specifically, I built two versions: one with a degree of 1 and another with a degree of 2. By building models with both degrees, we can assess whether capturing the interaction levels improves the model’s predictive performance. It is important to note that the target variable, ‘sleep\_efficiency’, is continuous, which means that it can take on any value within a specific range, unlike a classification problem where the target variable has discrete categories.

To evaluate the models’ effectiveness in predicting ‘sleep\_efficiency’, I calculated the Root Mean Squared Error (RMSE) for each model. Since RMSE measures the difference between predicted and actual values, a lower RMSE indicates a better fit for the continuous target variable. By comparing the RMSE of the Random Forest and MARS models for each question, we can determine which model performed better in predicting ‘sleep\_efficiency’.

Research Question 1

Figure 11 shows the R codes used to build the random forest model.

A computer screen shot of a computer code

Description automatically generated

Figure 11. Random Forest model for Research Question 1

Figure 12 shows the plot of random forest which confirms the error stabilised before 500 trees while Figure 13 shows the RMSE result of the model, which returned 0.1087677.

A graph of a tree

Description automatically generated

Figure 12. Random Forest model plot for Research Question 1

A close-up of a computer code

Description automatically generated

Figure 13. RMSE of random forest model for Research Question 1

Figure 14 shows the R codes used to build the two MARS model with degree 1 and degree 2, followed by their RMSE calculations of 0.1151711 and 0.1228908 respectively.

A computer code with blue text

Description automatically generated

Figure 14. MARS model and RMSE for Research Question 1

The best model according to the lowest RMSE is random forest for Research Question 1.

|  |  |  |
| --- | --- | --- |
| **Research Question 1 Performance** | | |
| **Models** | **RMSE** | **Best Model** |
| **Random Forest** | **0.1088** | 🗸 |
| **MARS (Degree 1)** | 0.1152 |  |
| **MARS (Degree 2)** | 0.1229 |  |

Research Question 2

Figure 15 shows the R codes used to build the random forest model.

A screenshot of a computer code

Description automatically generated

Figure 15. Random Forest model for Research Question 2

Figure 16 shows the plot of random forest which confirms the error stabilised before 500 trees while Figure 17 shows the RMSE result of the model, which returned 0.05503376.

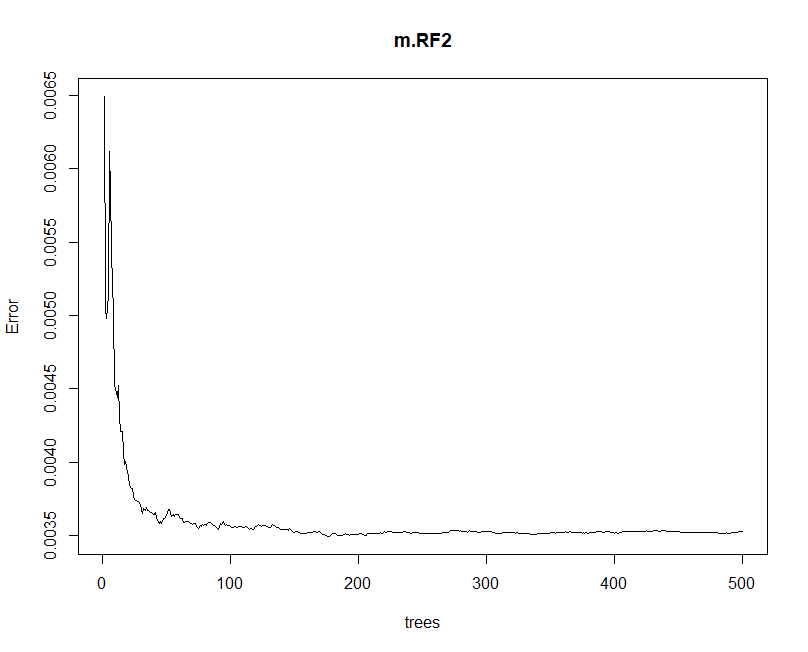


Figure 16. Random Forest model plot for Research Question 2

A close-up of a number

Description automatically generated

Figure 17. RMSE of random forest model for Research Question 2

Figure 18 shows the R codes used to build the two MARS model with degree 1 and degree 2, followed by their RMSE calculations of 0.570829 and 0.05692109 respectively.

A computer code with blue text

Description automatically generated

Figure 18. MARS model and RMSE for Research Question 2

The best model according to the lowest RMSE is random forest for Research Question 2.

|  |  |  |
| --- | --- | --- |
| **Research Question 1 Performance** | | |
| **Models** | **RMSE** | **Best Model** |
| **Random Forest** | **0.0550** | 🗸 |
| **MARS (Degree 1)** | 0.0571 |  |
| **MARS (Degree 2)** | 0.0569 |  |

# Answer to Q5:

Research Question 1

The analysis suggests that sleep efficiency can be predicted based on an individual’s lifestyle factors and habits, evident from the performance of the models. The performance evaluation of the models revealed that Random Forest achieved the lowest RMSE of 0.1088, which suggests its superior predictive accuracy as compared to the two MARS models with RMSE of 0.1152 (degree 1) and 0.1229 (degree 2). Based on the Random Forest predictive model analysis, I investigated the relationship between sleep efficiency and individual lifestyle factors and habits. We conducted variable importance ranking to identify the key predictors of sleep efficiency (Figure 19) and our findings indicated that ‘alcohol\_consumption’ had the highest %IncMSE, which shows how much the model accuracy would decrease should it be left out. This highlights its influence on sleep efficiency, followed by smoking status, exercise frequency, bedtime, age, and daily steps in descending order of importance. This highlights the significance of lifestyle choices and habits, particularly alcohol consumption and smoking status, in predicting sleep efficiency, while also acknowledging the impact of age. Overall, this analysis demonstrates an individual’s lifestyle choices have a measurable impact on sleep efficiency.

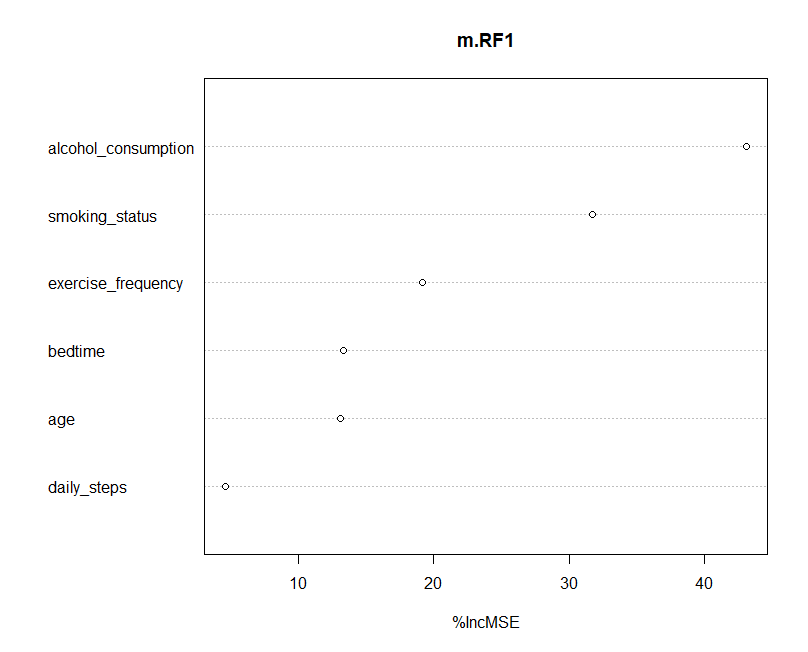


Figure 19. Variable Importance of RF Model 1

To supplement this, I zoomed in on the variable importance of the second top-performing model which is MARS with a degree of 1 (Figure 20). The results indicate that alcohol consumption and smoking status were the most important predictors, which complements the results from the random forest model.

A screenshot of a computer code

Description automatically generated

Figure 20. Variable Importance of MARS model 1 with degree = 1

Research Question 2

The analysis suggests that sleep efficiency can be predicted very well based solely on physiological sleep metrics. This is evident from the performance of the two models, Random Forest and MARS, on predicting sleep efficiency. The random forest achieved a very low RMSE of 0.055, indicating a highly accurate fit for the continuous target variable. Additionally, both MARS models performed comparably well, with RMSEs of 0.0571 (degree 1) and 0.0569 (degree 2).

When examining the variable importance in the random forest model (Figure 21), deep sleep percentage is the strongest predictor of sleep efficiency, followed by the number of awakenings experienced during sleep, and REM sleep percentage. These findings highlight the significant role of physiological sleep characteristics in determining overall sleep efficiency. Notably, this model achieved a lower RMSE compared to the model based on lifestyle factors in Research Question 1, suggesting that physiological sleep metrics might be more directly linked to sleep efficiency.

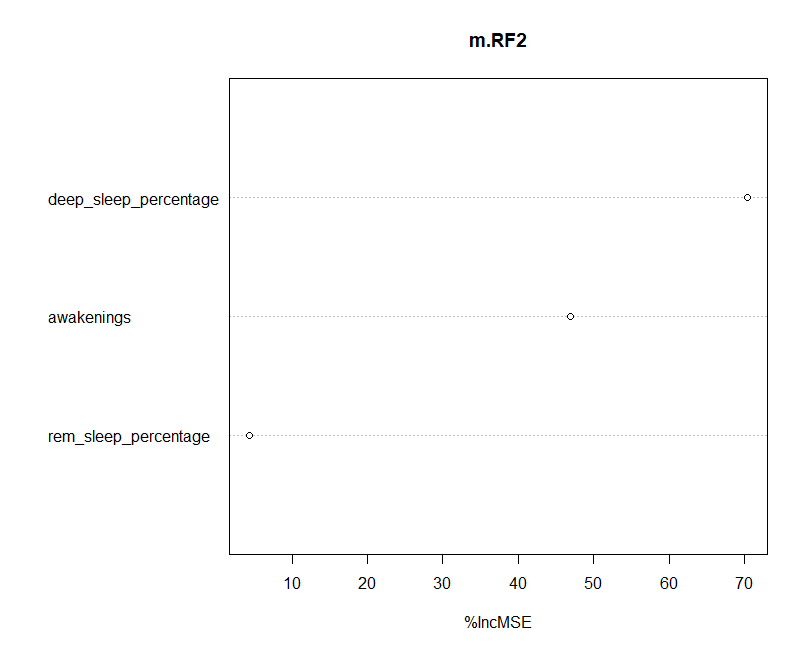


Figure 21. Variable Importance of RF Model 2

Similarly, I focused on the variable importance of the second top-performing model which is MARS with a degree of 2 (Figure 22). The results indicate that deep sleep percentage and awakenings were the most important predictors, which complements the results from the random forest model.

A close-up of a computer code

Description automatically generated

Figure 22. Variable Importance of MARS model 1 with degree = 2

# Answer to Q6:

This analysis explored the factors influencing sleep efficiency, a measure of how well an individual sleeps through the night. We investigated two key questions:

1. **Can lifestyle factors and habits predict sleep efficiency?**
2. **Can physiological sleep metrics predict sleep efficiency without considering lifestyle factors and habits?**

The answer to both questions is **yes.**

For lifestyle factors and habits, the analysis identified a clear connection to sleep efficiency. The random forest model achieved a good fit with an **RMSE of 0.1088**. Among these factors, **alcohol consumption** emerged as the most significant influence. This suggests that people who consumed more alcohol in the 4 hours prior to bedtime tended to have lower sleep efficiency. This was followed by factors like **smoking status, exercise frequency, and bedtime.** Conversely, variables like daily steps and age seem to have a less significant influence.

However, when we solely looked at **physiological sleep metrics**, the model achieved an even **higher level of accuracy** in predicting sleep efficiency. This data includes percentages of deep sleep and REM stages, as well as the number of awakenings during sleep. The random forest achieved a **RMSE of 0.055**, indicating **high prediction accuracy**. **Deep sleep percentage** was the strongest predictor, which meant that people with a higher percentage of deep sleep had better sleep efficiency, which verified the correlation analysis done during data exploration. Additionally, the number of **awakenings** and **REM sleep percentage** also played a role.

Overall, these findings suggest that both lifestyle choices and physiological factors influence sleep efficiency. Interestingly, **physiological sleep metrics seem to have a more direct impact** on sleep efficiency as compared to lifestyle factors. This highlights the importance of prioritising deep sleep and minimising awakenings during the night as these factors are crucial for optimal sleep efficiency.

# Answer to Q7:

Sleep is a fundamental physiological process that impacts one’s overall health and well-being. Since the 1970s, wearable technology has been used to assess sleep patterns in real-world conditions (Grandner et al., 2021). Currently, the wearables show promise in tracking when one is awake or asleep, but they need to improve accuracy in detecting sleep stages (Glazier, 2024). With scientifically proven studies done on the importance of sleep coupled with the growing adoption of sleep trackers, this is an opportunity to utilise machine learning models to enhance the population’s sleep health.

The model that is focused on physiological sleep metrics can be integrated into a personalised sleep tracking application which is then tied to a wearable sleep tracker to collect data on the different sleep stages (deep sleep, REM sleep, light sleep) and awakenings of the user. The application can then utilise the physiological sleep metrics model that is trained on a broader dataset to predict the user’s sleep efficiency based on their unique sleep patterns. By analysing factors like the user’s deep sleep percentage and awakenings, the app can also provide personalised recommendations. For instance, users with low deep sleep percentage might receive suggestions to engage in relaxation techniques before bed to aid a deeper sleep state, whereas for users with more frequent awakenings, the application can recommend strategies to create a better sleeping environment. With the ability to track the user’s progress and receive feedback from users, the model can learn and improve its personalised suggestions. This application would empower users to take charge of their sleep health with data-driven insights and actionable advice, and with better sleep health, it will be able to enhance the users’ overall well-being.

# References

Glazier, G. (2024, February 9). *New research evaluates accuracy of sleep trackers*. Sleep Foundation. https://www.sleepfoundation.org/sleep-news/new-research-evaluates-accuracy-of-sleep-trackers

Grandner, M. A., Lujan, M. R., & Ghani, S. B. (2021). Sleep-tracking technology in scientific research: Looking to the future. *Sleep*, *44*(5). https://doi.org/10.1093/sleep/zsab071

Hayes, A. (2022, January 10). *Stepwise regression: Definition, uses, example, and limitations*. Investopedia. https://www.investopedia.com/terms/s/stepwise-regression.asp