Answer to Q1:

FINDING 1 : Great Variability in Sleep Efficiency

Figure 1.1 shows the distribution of Sleep Efficiency using a histogram. It has a range of values from 0 to 1, with a concentration around 0.8 to 0.9. Overall, the histogram suggests that there is variability in sleep efficiency among individuals, which makes this variable a good candidate for prediction.

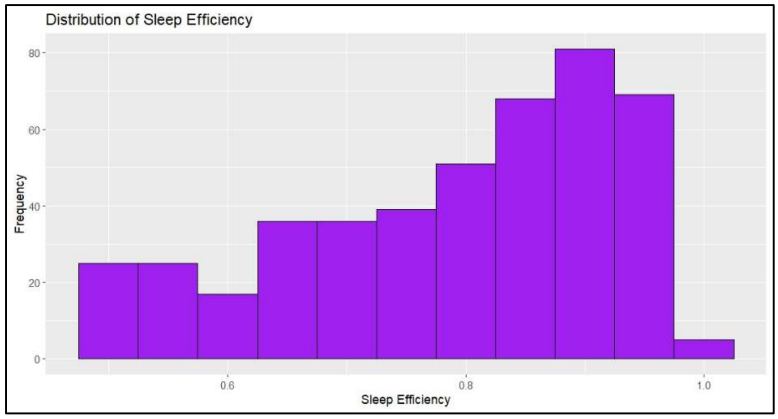


Figure 1.1

FINDING 2 : Measured Independently, Higher Caffeine Consumption and Exercise Frequency Correlate to Higher Sleep Efficiency

Figure 1.2 and figure 1.3 displays scatter plots of Caffeine Consumption against Sleep Efficiency and Exercise Frequency against Sleep Efficiency respectively. In both graphs, there seems to be an increasing trend line in which consuming more caffeine or exercising more often leads to more efficient sleep. Another observation is that for Caffeine Consumption, the data has very poor distribution, which hints that this variable may not be as influential (significant) in predicting the model.

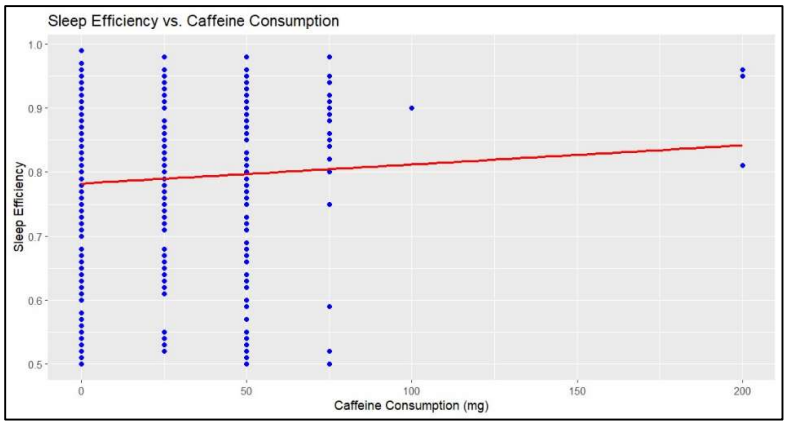


Figure 1.2

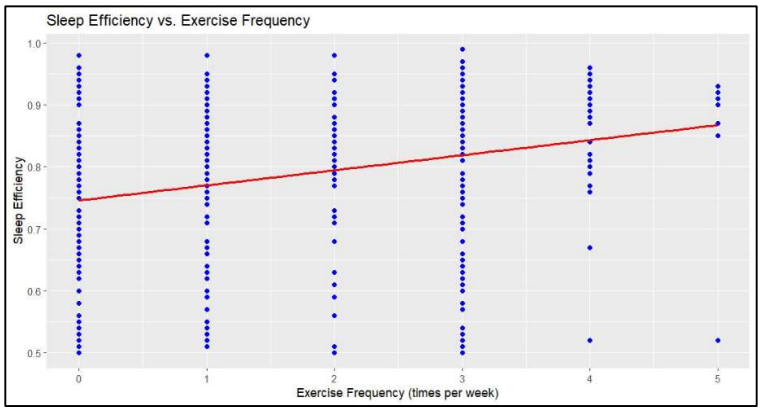


Figure 1.3

FINDING 3 : Measured Independently, Higher Alcohol Consumption and Awakenings Correlate to Lower Sleep Efficiency

Using the same scatter plot visualizations, figures 1.4 and 1.5 show distribution of Higher Alcohol Consumption and Awakenings against Sleep Efficiency. When analysed individually, an increase in each of the 2 variables seems to lead to **lower** sleep efficiency, which proves what is widely known in society. Statistically, both variables have better distribution than other variables in Finding 2, and they might (or might not) be more significant when used in a linear regression predictive model.

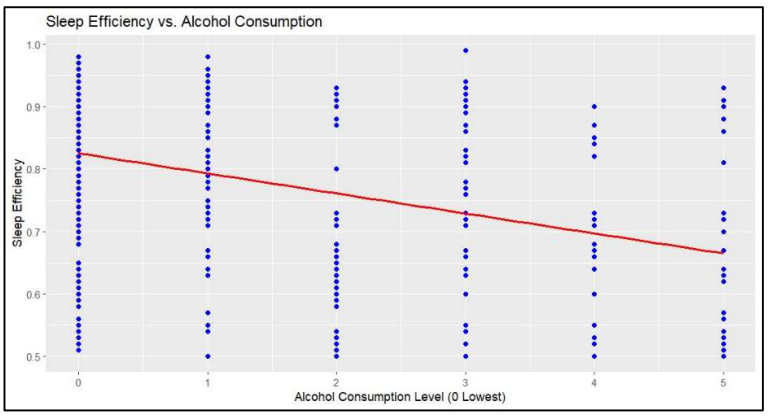


Figure 1.4

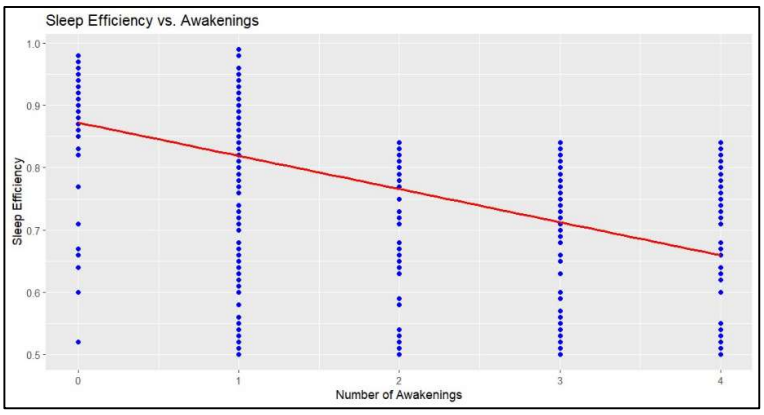


Figure 1.5

Answer to Q2:

From the analysis of all dataset features and insights regarding sleep, ***Sleep Efficiency*** **stands out as the best candidate for the target variable Y**.

Sleep Efficiency captures both the quantity and quality of sleep, providing a considerably reliable measure of **sleep health.** Unlike simply measuring total sleep duration, it **accounts for the proportion of time spent sleeping while in bed, offering a more comprehensive assessment of sleep quality**. Additionally, *Sleep Efficiency* is a clinically recognized metric in sleep studies, commonly used to diagnose and monitor sleep disorders. For instance, normal sleep efficiency is considered to be 80% or greater[1]. Furthermore, *Sleep Efficiency* integrates multiple dimensions of sleep into a single metric by considering both total sleep time and the efficiency of different sleep stages, such as REM (Rapid Eye Movement) percentages and deep sleep quantity. This integration ensures a balanced view of sleep health, capturing aspects that individual stage measurements might miss. If *Sleep Efficiency* is the target variable, models can be used to investigate factors/features that contribute most to its variability (as seen from the EDA section where this variable has a spread-out distribution).

Now that *Sleep Efficiency* is the predicted variable, a research question can then be formed and answered using analytic models: out of the many contributing variables,

What primary factors and pre-sleeping behaviors affect sleep productivity?

[1] https://www.sleepwa.com.au/interpreting-a-sleep-study/#:~:text=Normal%20sleep%20efficiency%20is%20considered,have%20sleep%20efficiencies%20above%2090%25.

Answer to Q3:

As mentioned in Q2 answers, the target variable Y will be *Sleep Efficiency*, a continuous variable. There are 2 types of models used; **Linear Regression model and CART model**. For both models, all variables except *Bed Time* and *Wakeup Time* are used to see which are labelled as significant or unimportant by each model automatically. Below is the analysis of each model, created using the same train set and then tested by the same test set as well.

<table>
<tr><th colspan="2">LINEAR REGRESSION MODEL</th></tr>
<tr><td>

```
> summary(model_v1)

Call:
lm(formula = `Sleep efficiency` ~ Awakenings + `Caffeine consumption` +
    `Alcohol consumption` + `Exercise frequency` + `Daily Steps` +
    `REM sleep percentage` + `Deep sleep percentage` + `Light sleep percentage` +
    Gender + `Smoking status` + Age + `Sleep duration`, data = dataTrain)

Residuals:
      Min        1Q    Median        3Q       Max
-0.178229 -0.042752  0.007766  0.042718  0.154069

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.432e-01  4.479e-02   7.662 1.82e-13 ***
Awakenings               -3.032e-02  2.588e-03 -11.717  < 2e-16 ***
`Caffeine consumption`    1.409e-04  1.085e-04   1.298 0.195107
`Alcohol consumption`    -5.761e-03  2.249e-03  -2.561 0.010847 *
`Exercise frequency`      4.792e-03  2.529e-03   1.895 0.058970 .
`Daily Steps`            -2.057e-06  1.620e-06  -1.270 0.204858
`REM sleep percentage`    7.141e-03  9.600e-04   7.438 7.99e-13 ***
`Deep sleep percentage`   5.716e-03  2.464e-04  23.199  < 2e-16 ***
`Light sleep percentage`         NA         NA      NA       NA
GenderMale                2.055e-03  7.256e-03   0.283 0.777220
`Smoking status`Yes      -4.506e-02  7.174e-03  -6.282 9.96e-10 ***
Age                       9.578e-04  2.519e-04   3.802 0.000169 ***
`Sleep duration`          1.676e-03  3.792e-03   0.442 0.658791
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06181 on 349 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.802
F-statistic: 133.6 on 11 and 349 DF,  p-value: < 2.2e-16

> vif(model_v1)
Error in vif.default(model_v1) :
  there are aliased coefficients in the model
```

</td><td>

LR Model V1

This first version of the linear regression model has an R-squared of **0.8081** which is quite ideal. However, the VIF function cannot be called, because **some variables are multicollinear with each other**.
More interpretations on this can be read under Q5 answers.

Furthermore, some variables are also found to be insignificant to the model (*Daily Steps, Caffeine Consumption, Sleep Duration, Gender, and Exercise Frequency*) and should be dropped along with one of the multicollinear variables.

</td></tr>
<tr><td>

```
> summary(model_v3)

Call:
lm(formula = `Sleep efficiency` ~ Awakenings + `Alcohol consumption` +
    `REM sleep percentage` + `Deep sleep percentage` + `Smoking status` +
    Age, data = dataTrain)

Residuals:
      Min        1Q    Median        3Q       Max
-0.179742 -0.041109  0.009606  0.042443  0.149380

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.3512384  0.0307865  11.409  < 2e-16 ***
Awakenings               -0.0319959  0.0024889 -12.855  < 2e-16 ***
`Alcohol consumption`    -0.0059089  0.0022220  -2.659  0.00819 **
`REM sleep percentage`    0.0072330  0.0009457   7.648 1.94e-13 ***
`Deep sleep percentage`   0.0057971  0.0002419  23.962  < 2e-16 ***
`Smoking status`Yes      -0.0450669  0.0071177  -6.332 7.35e-10 ***
Age                       0.0009710  0.0002435   3.988 8.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06198 on 354 degrees of freedom
Multiple R-squared:  0.8042,    Adjusted R-squared:  0.8009
F-statistic: 242.4 on 6 and 354 DF,  p-value: < 2.2e-16

> vif(model_v3)
        Awakenings   `Alcohol consumption`   `REM sleep percentage` `Deep sleep percentage`
`Smoking status`
          1.133441                1.213219                1.076912                1.406387
1.089490
               Age
          1.016250
```

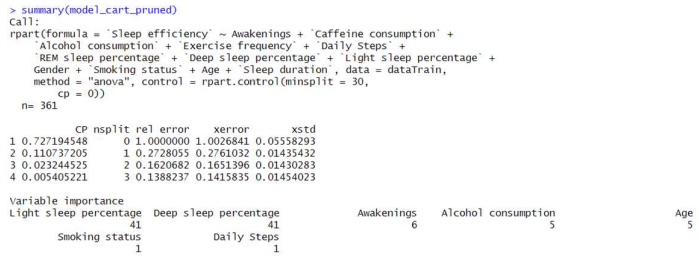
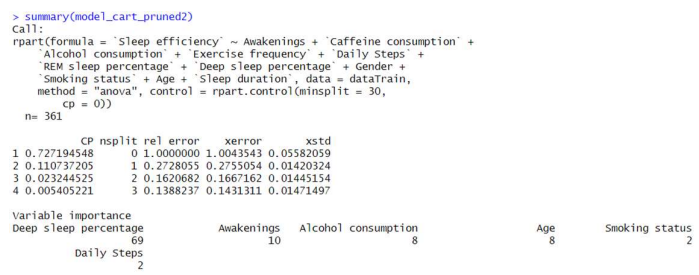
</td><td>

LR Final Model

This is the result **after some variables are dropped**, leaving only the most significant ones (seen by the ** and ***).

The R-squared error now is still ideal but slightly less than the previous model, being **0.8042**.
Here, there are no more multicollinear variables.

</td></tr>
</table>

CART MODEL	
	CART Model V1 This first CART model uses the exact same variables as the first version of the LR Model, with no care on which variables are significant, multicollinear, and so on.
	CART Model V2 This model is created for comparison, where the only difference from CART Model V1 is that this **includes *Deep Sleep Percentage* but excludes *Light Sleep Percentage*** (prevents multicollinearity).

TEST RESULTS COMPARISON		
LR Final Model	CART Model V1	CART Model V2
> head(linreg_comparison, 10) Actual Predicted 1 0.89 0.9085366 2 0.98 0.9201455 3 0.89 0.8606349 4 0.91 0.9173167 5 0.84 0.7975803 6 0.65 0.5888162 7 0.87 0.8635261 8 0.81 0.8252884 9 0.84 0.8407125 10 0.93 0.8745899	> head(cart_comparison, 10) Actual Predicted 1 0.89 0.8905114 2 0.98 0.8905114 3 0.89 0.8905114 4 0.91 0.8905114 5 0.84 0.7751163 6 0.65 0.6345238 7 0.87 0.8905114 8 0.81 0.8905114 9 0.84 0.7751163 10 0.93 0.8905114	> head(cart_comparison2, 10) Actual Predicted 1 0.89 0.8905114 2 0.98 0.8905114 3 0.89 0.8905114 4 0.91 0.8905114 5 0.84 0.7751163 6 0.65 0.6345238 7 0.87 0.8905114 8 0.81 0.8905114 9 0.84 0.7751163 10 0.93 0.8905114
RMSE : 0.05815137	**RMSE : 0.05489136**	**RMSE : 0.05489136**

Evaluation after testing shows that the **CART models perform better than the LR model**, though not by a dramatic amount. Overall, it can be said that both models perform quite well, as RMSE is expected to be as close to 0 as possible. The *head()* function shows first 10 columns from test data of the *actual Sleep Efficiency* values and their respective predictions for each model.

Additional comparisons between the 2 CART models can be read in Q5 answers.

Answer to Q4:

Research Question :
What primary factors and pre-sleeping behaviors affect sleep productivity?

The **Linear Regression Model answers** that based on the t-values, the primary factors or pre-sleeping behaviors affecting sleep productivity (efficiency) by decreasing importance are **Deep Sleep Percentage, Awakenings, Smoking Status Yes, REM Sleep Percentage, Age, and Alcohol Consumption**.

Meanwhile, the two **CART Models consider** these variables important in decreasing order: **Light/Deep Sleep Percentage, Awakenings, Alcohol Consumption, Age, Smoking Status, and Daily Steps**.

- In CART Model V1 where it considers both Light Sleep Percentage and Deep Sleep Percentage, each of variables have the same importance points of 41.
- In CART Model V2 where there is no Light Sleep Percentage and only Deep Sleep Percentage exists, the variable importance of Deep Sleep Percentage is higher (69). This decision skews the 'importance' ratios of other variables in Model V2, but the order of rankings of significant variables remains the same.

Different models seem to provide different answers to the research question. Nevertheless, the variables that are **deemed important in both models** are: **Deep Sleep Percentage, Awakenings, Smoking Status, Age, and Alcohol Consumption**.

It can be said that these 5 factors have definite significant effect on how efficient and productive a person's sleep is, regardless of their bed time and wake up time.

Meanwhile, the insignificant factors that are not considered by both models are also worth recognizing: Caffeine Consumption, Gender, Sleep Duration, and Exercise Frequency are not believed to influence a person's quality of sleep that much, as they don't appear in both LR Final Model and the CART models.

Answer to Q5:

Additional Finding 1: Light Sleep Percentage is multicollinear with Deep Sleep Percentage.

Light Sleep Percentage is seen to show *N/A* on all table, which suggests that it is multicollinear with another variable. The likely candidate is *Deep Sleep Percentage,* since the two combined accumulate to 100%, providing the same 'meaning' and thus perfectly multicollinear. To handle this, either one of the variables must be removed from the model (and once this is done, VIF can be measured).

Additional Finding 2: REM Sleep Percentage Correlation with Sleep Efficiency is Counterintuitive.

The Coefficient for REM (Rapid Eye Movement) Sleep Percentage in LR Final Model is 0.007230, and its significance (P-value) is 1.94e-13. The direct interpretation is that REM Sleep Percentage becomes a positive predictor of Sleep Efficiency, where a 1% increase in REM sleep percentage results in an increase in sleep efficiency by approximately 0.007.
This is an odd finding that may need further research, since in the real world, the opposite is true: increasing REM during sleep leads to decreasing sleep efficiency. There might be some domain knowledge required to address this case.

Additional Finding 3: Presence of multicollinear variables has little to no effect in CART Model performance.

This is perhaps one of the most interesting findings out of all. The CART Models, as explained before, are different in the way that the first model contains the multicollinear variables while the second model does not. Yet not only do they have the same RMSE, their predicted values for each test data are also the same. This shows that CART handles multicollinearity extremely well, without the need for manual elimination and analysis of variable-by-variable significance like linear regression models. Seen in Figure 5.1, both CART models have the same exact tree plot after their most optimal tree is pruned.

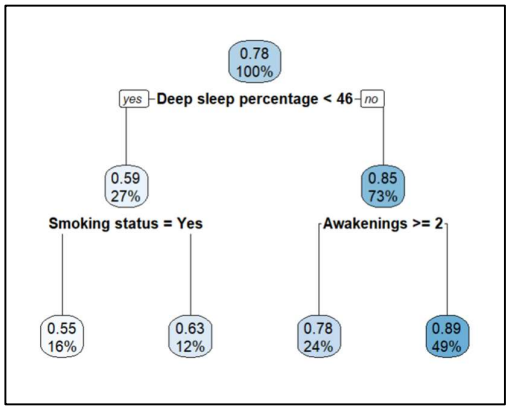


Figure 5.1

Answer to Q6:

The predictive model created for *Sleep Efficiency*, which incorporates factors such as Deep Sleep Percentage, Awakenings, Smoking Status, Age, and Alcohol Consumption, holds significant potential in **improving the quality of life in nursing homes for the elderly**.

In environments where elderly residents are continuously monitored for better care, the model offers an opportunity to gain insights on factors that influence their sleep efficiency. There can be personalized interventions aimed at maintaining sleep quality based on each residents' lifestyle factors: whether they smoke or not, whether they enjoy alcohol, and so on.

In practical terms, the model can guide healthcare professionals in balancing lifestyle choices and health conditions to optimize sleep efficiency. Below are some of the possible applications :

- The model can provide recommendations on the permissible levels of alcohol consumption for elderly residents, taking into account their smoking status and age. This gives some degree of freedom on allowances for alcohol as long as it does not adversely affect sleep quality.
- By monitoring and adjusting for factors such as deep sleep percentage and the number of awakenings, healthcare providers can implement activity plans to improve sleep patterns. This proactive approach not only enhances the overall well-being of residents but also potentially reduces the risk of sleep-related health issues, thereby contributing to a higher standard of care in nursing homes.