Aalto University
School of Science

CS-C4100 - Digital Health and Human Behavior

Sleep analysis

Image credit: Design by Freepik Image by macrovector on Freepik

Table of content:

1. Introduction

Sleep is a basic need in every individual's life which heavily influences both their physical and mental health. During my time as a student, I have witnessed my friends as well as myself struggling with maintaining high sleep quality, yet, unable to effectively improve our rest. If the factors correlated to quality of sleep could be identified, it would be easier to enhance or mitigate them accordingly.

There are various factors that may contribute to an individual's quality of sleep. It has been suggested that a high stress level is associated with bad sleep quality (Alotaibi, Alosaimi, Alajlan & Bin Abdulrahman, 2020). In this research, the researchers administered a questionnaire to the participants to assess their stress level in general. There are not yet any studies that specifically look at whether encountering a stressful day would affect that night's sleep. This seems to be worthy of investigation.

It also has been studied that regular physical activity improves sleep quality (Alnawwar, Alraddadi, Algethmi, Salem, Salem & Alharbi, 2023). Even on a smaller scale, more active days are also correlated with longer and better sleep at night (Sullivan Bisson, Robinson & Lachman, 2019). Hence, it would be useful to learn if light activities such as walking would have an impact on sleep. This result can then be utilised to help recommend how much an individual should exercise for better sleep.

Caffeine is a widely consumed stimulant that is a part of many people's morning routine in the form of coffee. It has been studied that there caffeine consumption itself has no impact on sleep quality (Del Brutto, Mera, Zambrano & Castillo, 2016), but rather, it is found that tiredness in the morning generally leads to high caffeine usage, where tiredness in the morning is associated with poor sleep quality (Snel & Lorist, 2011). However, the observation of the lack of correlation between caffeine consumption and sleep quality in the research by Del Brutto et al. was made in a rural environment where the participants are not affected by other variables such as light or noise pollution. This means that it would be interesting to use another set of data to assess the correlation.

It has been pointed out that students with irregular bedtime have poor sleep quality (Kang & Chen, 2009). Hence, it would be useful to learn the degree of correlation between these factors, and examine how much sleep time irregularity can negatively affect sleep.

Therefore, in this paper, I will be embarking on identifying and analysing the various possible correlations between different factors and sleep quality. After that, I will be exploring if sleep quality can be predicted using a subject's personal data via the two

machine learning methods: linear regression and logistic regression, and try to see in what way can the models be improved to yield better results.

In terms of analysis, I have created heatmaps of correlation and pair plot, as well as looked at the distribution in histograms to see how some factors affect sleep quality. For the machine learning methods, I have selected quality as the label, and the subject's personal data as the features. The logistic regression method can only be used for binary classification; thus, I have classified the data for the label into two categories of "Better sleep quality" and "Worse sleep quality". In the linear regression method, the value of mean squared error is used to judge how well the model performs, and for the logistic regression method, accuracy score and precision score are used.

The key findings of this paper reveals that only time asleep and time in bed have a fairly strong correlation to sleep quality, and that both of the models mentioned above yield relatively good results in predicting sleep quality. Both linear regression and logistic regression performs better when there are more features used in fitting, even if there are significantly fewer data points for training.

2. Problem Formulation

In this paper, I will be trying to explore associations between various factors that take place during the day and sleep quality at night, as well as the degrees of these associations. These factors include:

- Whether or not stressful situations were encountered;
- Number of steps taken,
- The level of caffeine intake;
- Regularity of sleep;
- The movement per hour during sleep;
- Whether or not a wake up alarm was set;
- Whether or not the individual snores during sleep.

Furthermore, I also aim to see if I can utilise these factors in the prediction of sleep quality using machine learning methods. After that, I will be assessing which methods would be better suited for this task, and how the model can be improved on to yield more accurate results.

By exploring all of these associations using various analysis techniques, I can learn more about factors that affect sleep quality, and gain some insights on possible ways to get better sleep. Moreover, using different techniques will allow me to assess their suitability with different data types and data distributions.

The assessment of the suitability of machine learning methods in predictions can open the possibility to predict a person's general quality of sleep using the personal information of their daily habits. This can potentially be applied in the wellness industry to diagnose and give clients suggestions more easily.

3. Dataset Description

For this project, I am utilising a dataset uploaded on Kaggle where the data was collected from Northcube's Sleep Cycle application on iOS between 2014 and 2022 for an individual user (Diotte, 2022). The data is actively collected via the user's inputs into the application, as well collected passively using the user's own device (Sleep Cycle, 2023). This utilises the device's functions such as the accelerometer, microphone, camera, and location. Therefore, each data point represents daily observations of routine, environment, and bodily activities for a single subject, which was collected both actively and passively, as well as some derived data such as "Sleep quality".

This dataset is split into 2 dataframes, one containing data points from between 2014 and 2018 with 887 entries, while the other contains data from 2018 to 2022 with 921 entries. Altogether, this dataset has 1808 data points.

The 2 dataframes for the dataset have different columns. The tables below show the column names, description and data types:

Dataframe with data collected from 2014 - 2018			
No.	**Column**	**Description**	**Data type**
1	Start	time the subject started sleeping [DateTime]	Continuous
2	End	time the subject stopped sleeping [DateTime]	Continuous
3	Sleep quality	[%]	Continuous
4	Time in bed	[hours and minutes]	Continuous
5	Wake up	Mood upon waking up [:) (Good), :| (Not good)]	Binary
6	Sleep Notes	[Drank coffee, Drank tea, Worked out, Ate late, Stressful day]	Categorical
7	Heart rate	[bpm]	Continuous
8	Activity (steps)	[steps]	Continuous

Table 1: Details on the 2014-2018 dataframes columns (8 columns)

Dataframe with data collected from 2018 - 2022			
No.	**Column**	**Description**	**Data type**
1	Start	time the subject started sleeping [DateTime]	Continuous
2	End	time the subject stopped sleeping [DateTime]	Continuous
3	Sleep Quality	[%]	Continuous
4	Time in bed (seconds)	[seconds]	Continuous
5	Mood	Completely missing, no available data	N/A
6	Notes	[Coffee, Tea, Worked out, Ate late, Kathryn's, Alcohol]	Categorical
7	Heart rate (bpm)	[bpm]	Continuous
8	Steps	[steps]	Continuous
9	Regularity	Regularity of sleep [%]	Continuous
10	Alarm mode	[Normal, No alarm]	Binary
11	Air Pressure (Pa)	[Pa]	Continuous
12	City	Name of the city where the individual spent the night	Categorical
13	Movements per hour	Movements during sleep	Continuous
14	Time asleep (seconds)	[seconds]	Continuous
15	Time before sleep (seconds)	[seconds]	Continuous
16	Window start	Unclear [DateTime]	Continuous
17	Window stop	Unclear [DateTime]	Continuous
18	Did snore	[true, false]	Binary
19	Snore time	[seconds]	Continuous
20	Weather temperature (℃)	[℃]	Continuous
21	Weather type	Type of weather during the day	Categorical

Table 2: Details on the 2018-2022 dataframes columns (21 columns)

For the data frame 1, the columns “Start”, “Time in bed” and “Sleep Notes” are used as features, and “Sleep quality” as the label. For data frame 2, the columns “Start”,

“Time in bed (seconds)”, “Time before sleep (seconds)”, “Time asleep (seconds)”, “Movements per hour”, “Alarm mode”, “Steps”, and “Regularity” are used as features, and “Sleep quality” as the label.

This dataset does not have a very good usability and has required quite a lot of data pre-processing. To decide on what preliminary data processing is needed, I have checked for duplicates as well as the number of missing values for each column of each data frame. For both of the data frames, I will be dropping columns where there are hundreds of missing values, since each of the data frames only has less than 1000 data points. Removing the null data is necessary, as not doing so will lead to incorrect results, as well as preventing some visualisation tools from working properly.

Next, upon closer inspection for data frame 2, I have realised that there are inputs with the value “0” under the “Air Pressure (Pa)” column. The normal atmospheric air pressure at sea level is about 101000 Pascal, and having the atmospheric pressure of 0 Pascal is equivalent to being more than a 100 kilometres above the sea level (Stull, 2019). Since commercial aircrafts only fly up to the altitude of around 12.8 kilometres at the maximum, it is virtually impossible for the air pressure where the subject stays to be 0 Pascal, unless the subject is an astronaut living in a space station. Nevertheless, such an important detail would be highly unlikely to be omitted from the data source. Therefore, I replaced these “0” inputs with the null value “NaN”. After replacement, I checked for the number of missing data in the “Air Pressure (Pa)” column. I have decided to drop this column as well since there is too much missing data. For the columns “Window start” and “Window stop” in data frame 2, there is no information on their significance. Thus, I have also dropped these columns. After dropping these columns for both dataset, there are no more missing inputs.

In dataframe 2, for the "Did snore" column, there are discrepancies with the "Snore time" column. Even when the snore time is registered as 0.0, the "Did snore" value can still be True. I will be changing the "Did snore" value into False if the snore time is 0.0.

Since the values of “Sleep quality” were entered as a string in the data frames, I have written a function to remove the “%” sign that comes with each value and turn them into integer values. This is such that I can use data visualisation tools on them.

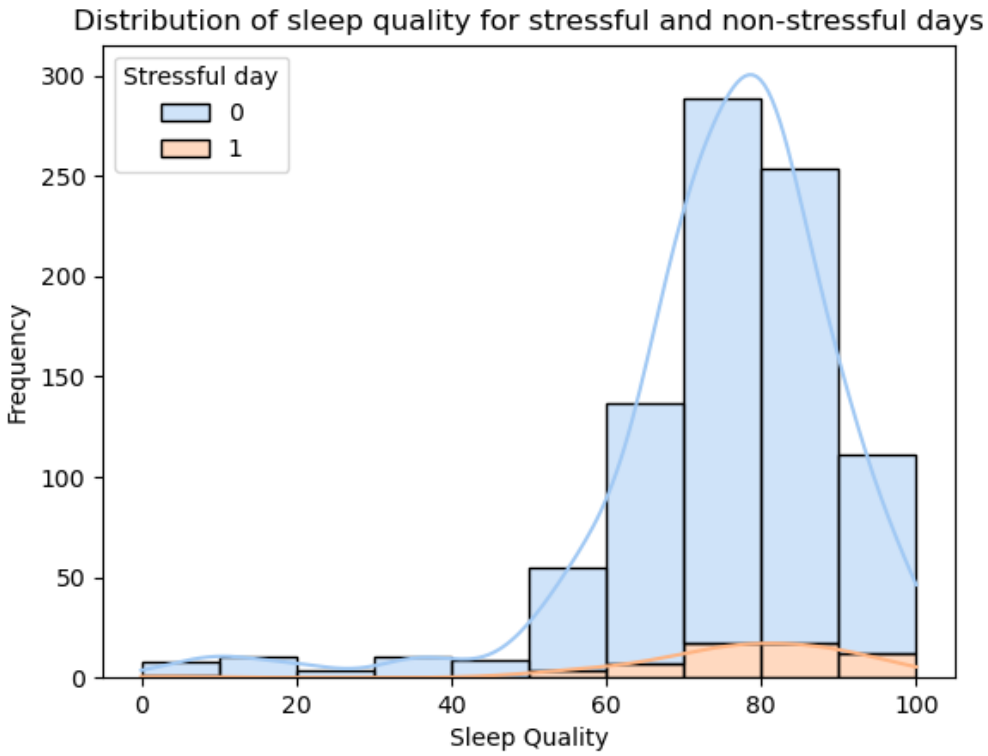
Next, as both data frames have some common columns, it is possible to concatenate the data frames vertically to yield a bigger data frame with more data points. Before concatenation, I renamed some columns such that the corresponding ones share the same names across the data frames. These common columns are "Start", "End", "Sleep quality", "Time in bed", and "Steps". After this process, I obtained a larger

data frame with 1808 entries. I will be referring to this data frame as “the concatenated data frame” in other instances of mention.

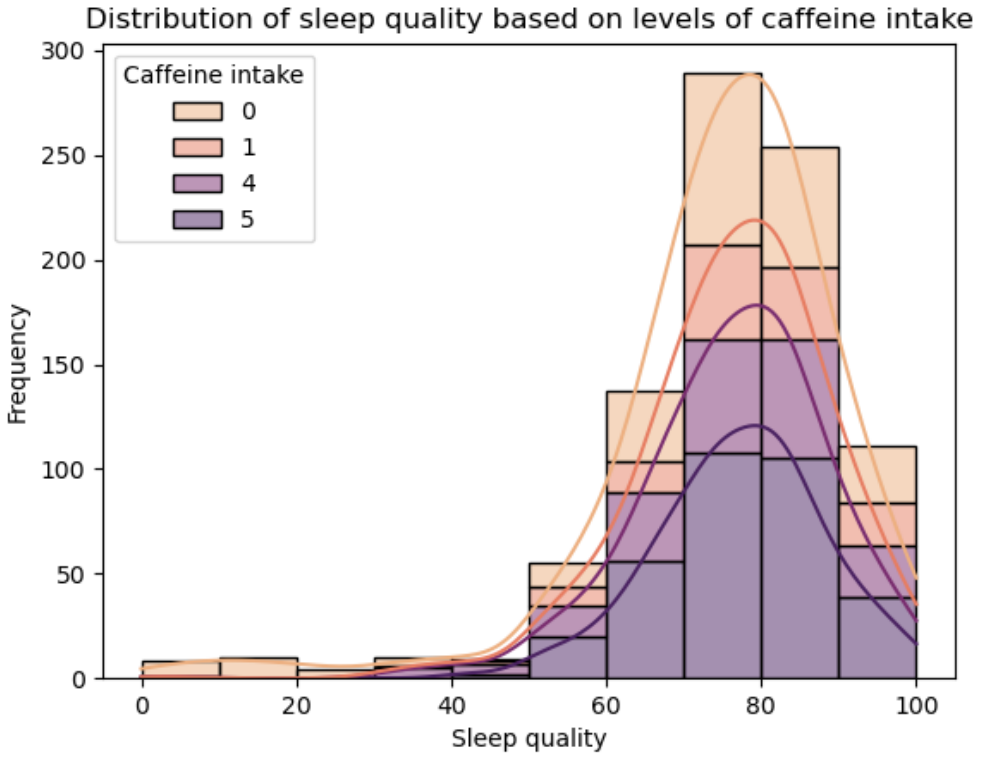
4. Methods

First, I will be presenting the observations I have made from analysing the data frame 1. For this data frame, I have decided to focus on “Sleep Notes” and its possible correlation to the quality of sleep. Under this column, there are two factors that I deem as meaningful for analysis. It would be interesting to see how the level of caffeine intake, or if the day was stressful, could have an impact on sleep quality. I will create a new dataframe from dataframe 1 where the string values of the column "Sleep Notes" are processed such that a new column "Stressful day" which has binary inputs can be created (where the value “1” signifies a stressful day, and the value “0” otherwise); as well as a new column "Caffeine intake" which arbitrarily quantify amount of caffeine taken during the day. Coffee typically contains 40 milligrams of caffeine per 100 grams portion (USDA, 2004), while tea typically contains 11 milligrams of caffeine per 100 grams portion (USDA, 2005). Hence, to arbitrarily quantify the amount of caffeine intake per day for the subject, I will assign tea the value "1" and coffee the value "4". This means that on days where the subject drinks both coffee and tea, the value of the “Caffeine intake” column would be “5”.

After this, I have plotted stacked histograms with 10 bins, and kernel density estimates which show the estimates of the data distribution. The stacked bars represent the distribution of sleep quality within the different categories. In the first visualisation below, the categories are “Stressful day” (equivalent to “1”) and “Non-Stressful day” (equivalent to “0”) .



In the second visualisation, the categories are the amount of caffeine taken during the day, which are arbitrarily “1”, “4”, or “5”.

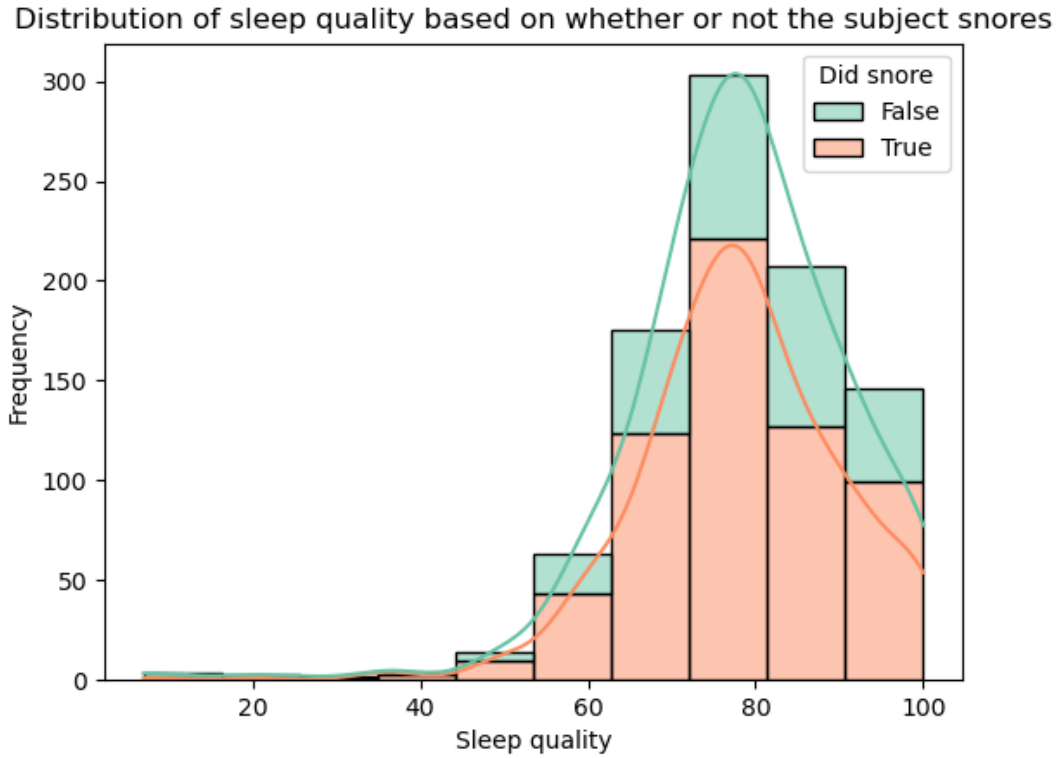


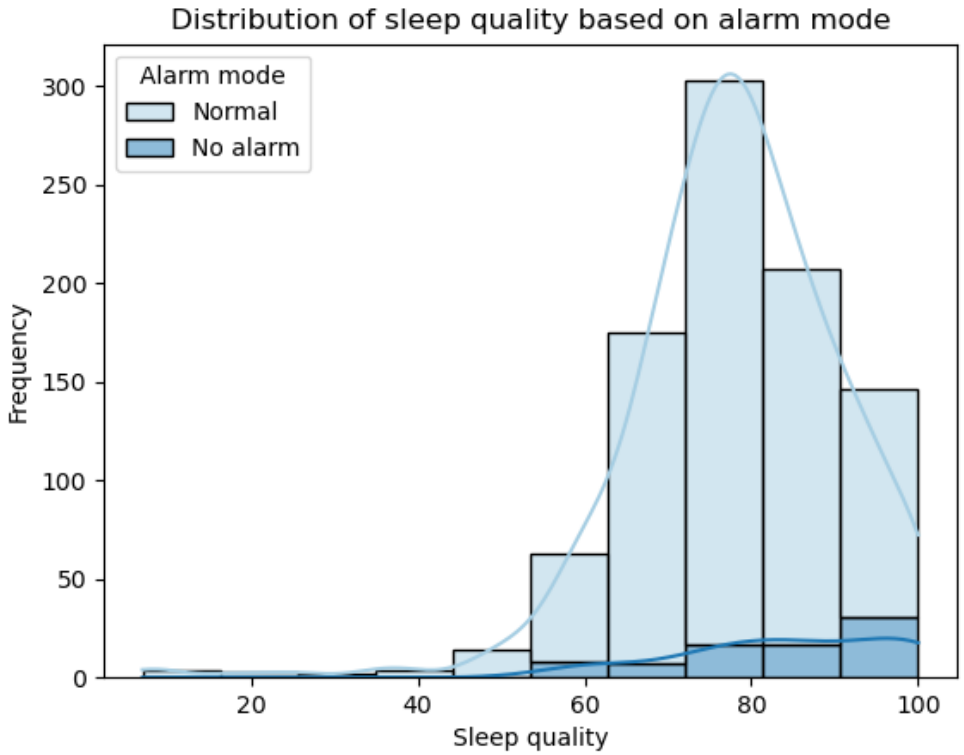
From the stacked histograms above, we see that the distribution of sleep quality follows the normal distribution and is left skewed, as the tail of the plot is dragged

towards the lower end of the x-axis. Looking at the kernel density estimates for both stressful days and caffeine intake, it is clear that their distribution all follow the distribution of sleep quality. This means that the factor of “whether or not the subject encountered some stressful situation during the day” and “caffeine intake” have no correlation to sleep quality.

Secondly, I analysed data frame 2. I have created similar stacked histograms which show the distribution of sleep quality based on whether or not the subject snores, and if they have set an alarm.

For the distribution of sleep quality on whether or not the subject snores below, we can see that the sleep quality for “the night the subject snores” and for “the night the subject does not” follow very similar distribution, suggesting that there is no correlation between snoring and quality of sleep for the subject.



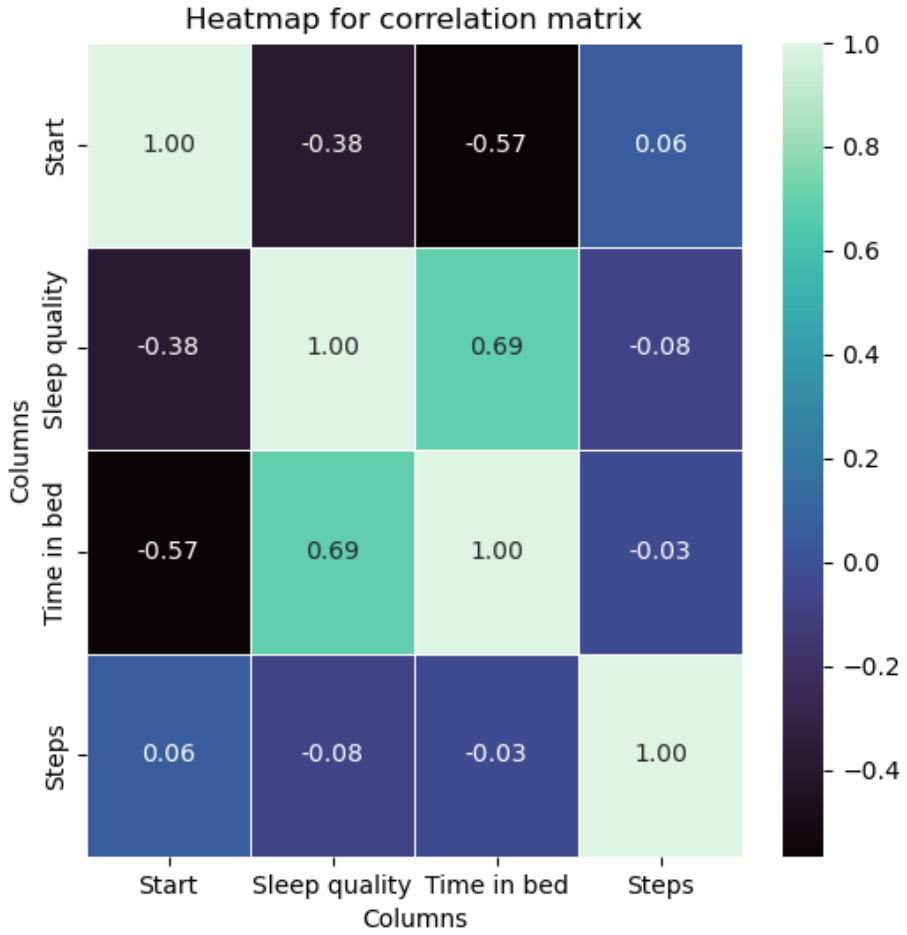


From the distribution of the sleep quality based on alarm mode, we can see that the distribution for sleep quality with no alarm does not follow the normal distribution and is more left skewed than for sleep quality with alarm. This suggests that the subject generally enjoys better sleep quality when there is no alarm.

Finally and most importantly, I moved on to looking at the concatenated data frame and implementing the machine learning methods. Before I could do any meaningful analysis on the data, I had to do a bit more of data processing. I converted the values in the column “Start” into pandas’ Date Time format, then removed the date and only kept the hours and minutes as a float value with “hours” as the unit. For example, “2014-12-29 22:57:49” is converted into “22.96361111”. The values in the column "End" can be calculated with the values in the columns "Start" and "Time in bed", so I dropped this column.

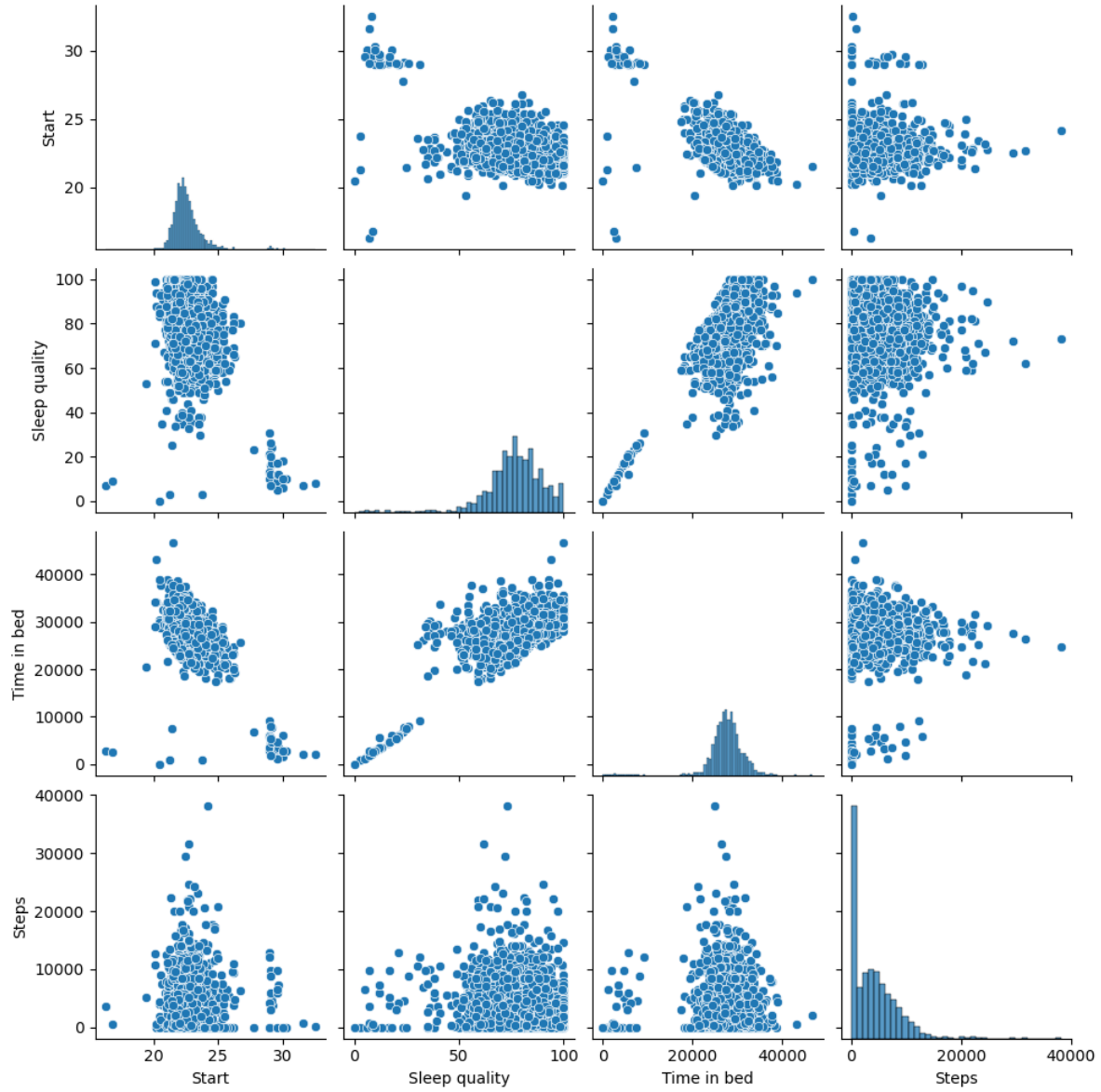
Furthermore, for the column "Start", the time is recorded in the 24-hour format; therefore, if the subject goes to sleep at 1am, the value is registered as 1. However, this can cause the plotting to be incoherent. Hence I will be changing the values such that 1am is registered as 25, 2am is registered as 26, and so on all the way until 11am.

To select the features that will be used for the machine learning models, it is necessary to look at the relationship between them first. I have used seaborn’s heatmap function to create a heatmap for the correlation matrix.



In order to reduce noise for the machine learning models, I have decided that only the personal data whose correlation with “Sleep quality” is larger or equal to 0.05, or smaller or equal to -0.05 are used. Thus, for both of the models, the label would be “Sleep quality”, while the features are “Time in bed”, “Start”, and “Steps”.

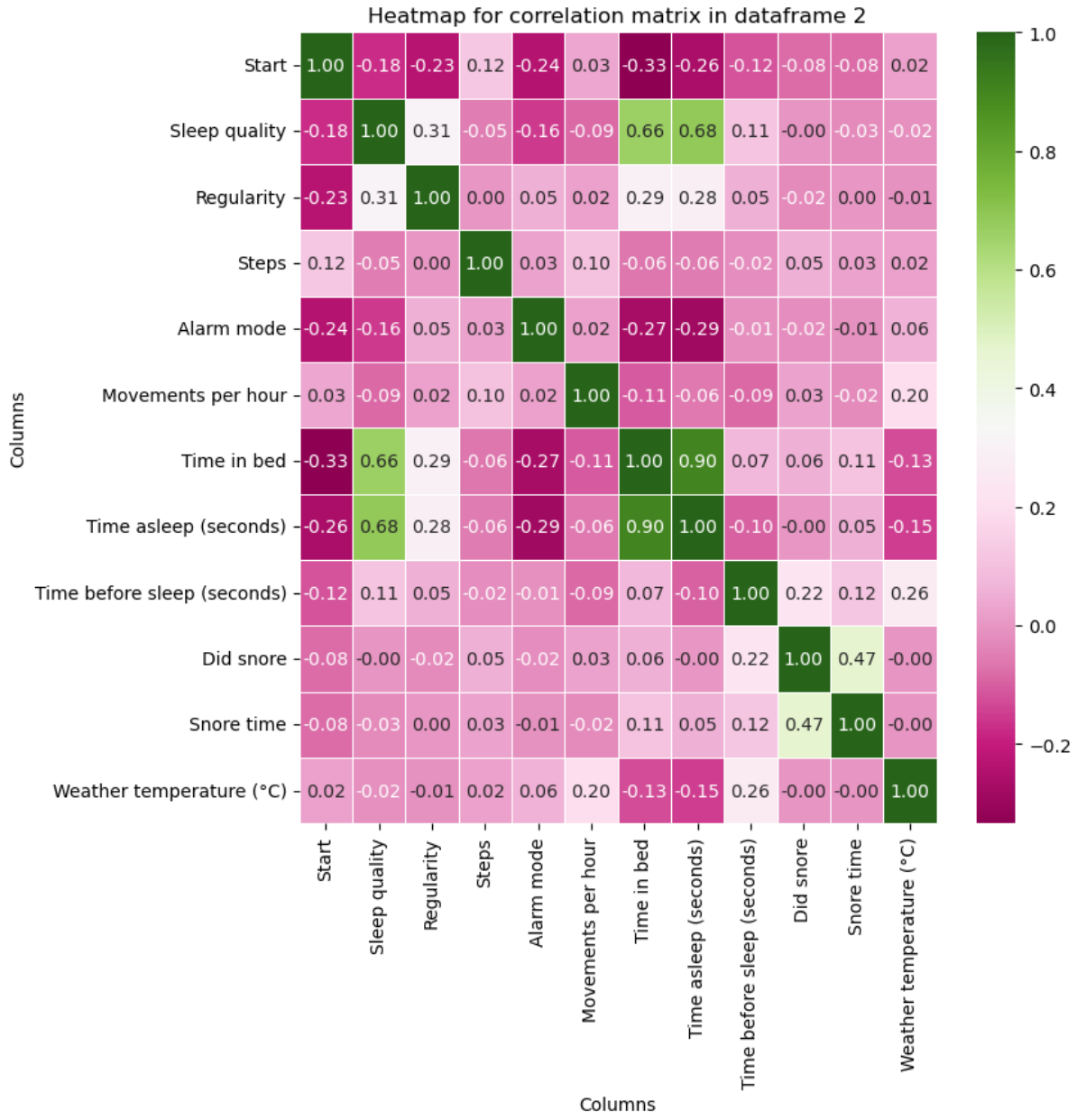
Let us also take a look at the scatter plots of all the columns in this dataframe.



For our first model, I have chosen the linear regression method as this method is suitable for predicting a continuous outcome in the form of a numeric value. In this case, one of the features - “Time in bed” - expresses a clear linear relationship with sleep quality as shown above, and linear regression is very efficient when there is a linear relationship between the label and features (Jung, 2022). I started the process with importing the linear regression model from sci-kit learn. Then, I initialised this model and assigned the label to the series y, and the features to the data frame X. Using the “fit” function of the linear regression model, our model is trained using X and y. After the model is trained, I used the “predict” function and passed X as the variable to obtain the predicted value for y - namely “y-pred”. Next, I created a scatter plot of the predicted value of y against the actual value of y with the line where y_pred = y to see how well linear regression performs. Lastly, I calculated the mean squared error to assess the performance of the model.

For the second model, I have chosen logistic regression. This is because in this case, many times, there is not really a need to predict the numerical value of sleep quality, but whether or not the sleep quality is “good”. This calls for binary classification, which the logistic regression model performs well. Logistic regression is also highly comprehensible and easily implemented, with high efficiency in training (Jung, 2022). To implement this model, I first categorised the values under “Sleep quality” into worse sleep, represented by the value “0”; and better sleep, represented by the value “1”. In terms of categorisation, I used the median sleep quality as the threshold. Sleep quality with the numerical value smaller than that of the median is assigned to “0”, while the sleep quality larger or equal to the median is assigned to “1”. I started the process with importing the logistic regression model, the function “train_test_split”, as well as the functions to calculate the accuracy score and the precision score from sci-kit learn. Then, I initialised this model and assigned the label to the series y, and the features to the data frame X. The data is then split into the training and testing sets using the “train_test_split” function, where the size of the testing set is 20% of the whole dataset. Since this dataset is big enough, “train_test_split” will generate the training and testing sets of ample size, and avoid over-fitting. After that, using the “fit” function, the model is fitted to the training data. Prediction is then made on the test data. I ended the process with calculating the model's accuracy and precision. Moreover, I have also created a confusion matrix to assess the model performance.

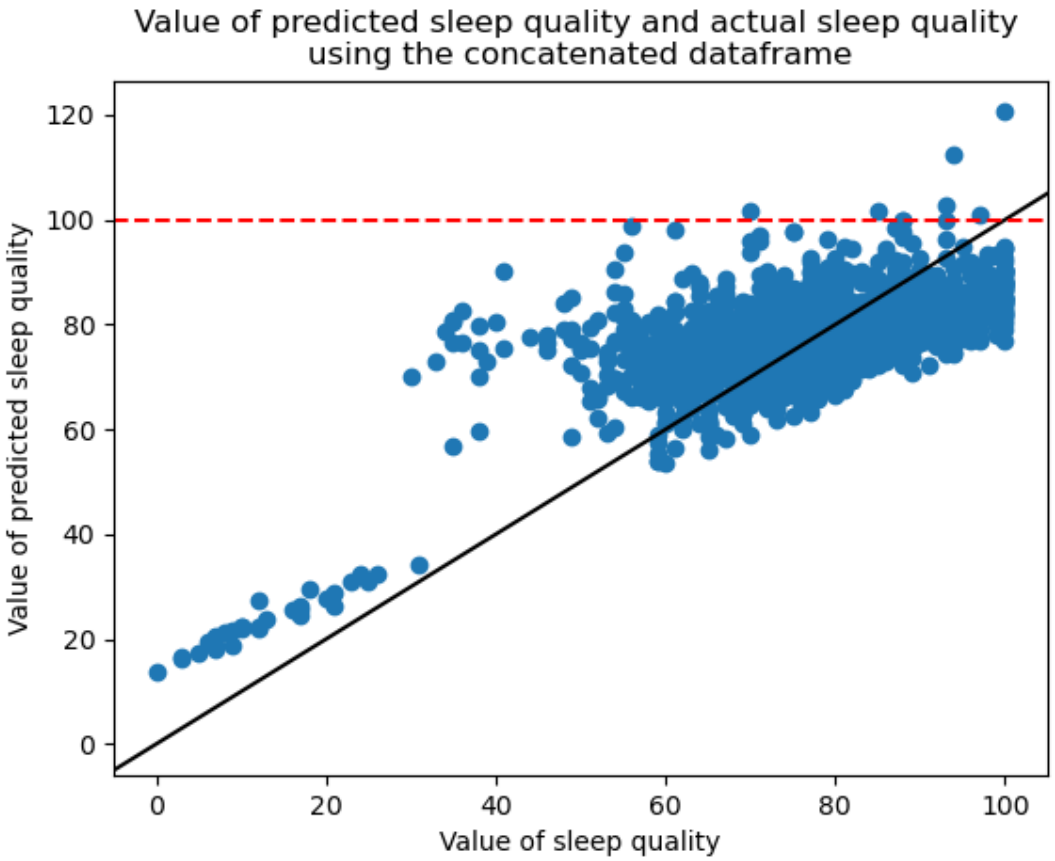
In order to learn more about the suitability of these machine learning methods, I have also implemented both of them on the data frame 2 using “Sleep quality” as the label while using “Start”, “Time in bed (seconds)”, “Time before sleep (seconds)”, “Time asleep (seconds)”, “Movements per hour”, “Alarm mode”, “Steps”, and “Regularity” as features. These features are chosen as their correlation to sleep quality is larger or equal to 0.05, or smaller or equal to -0.05, as seen below.



5. Results

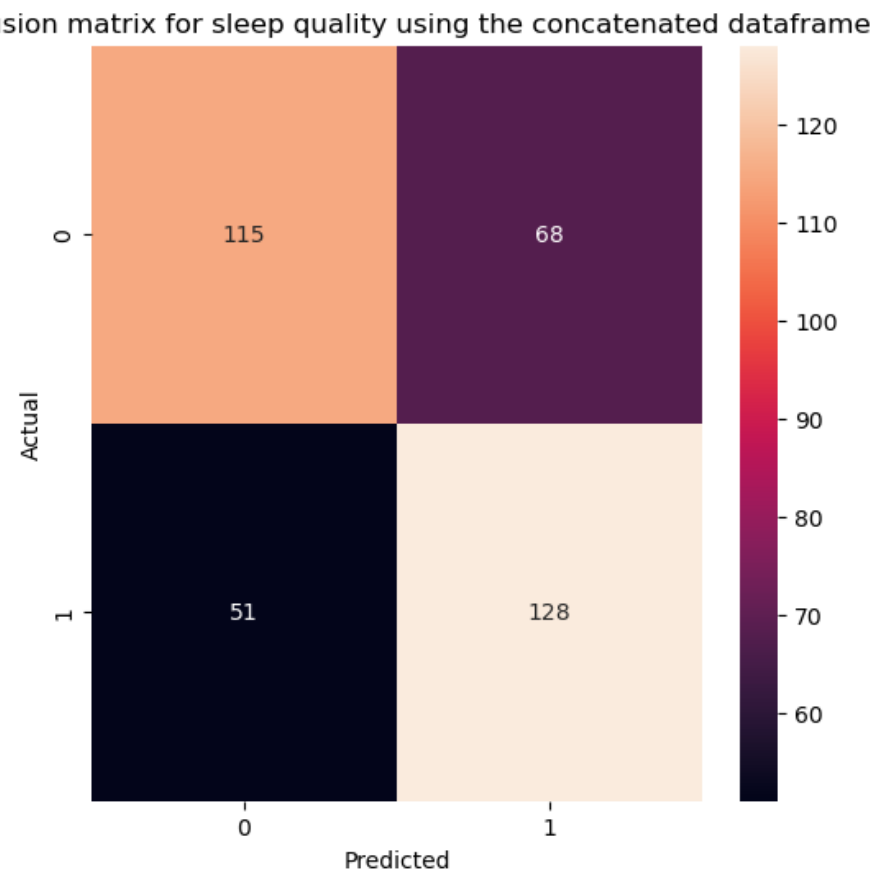
From the analysis above, we can see that the subject has a generally good sleep quality. The factors that have the strongest correlation to sleep quality are “Time in bed” and “Time asleep”, while other factors such as “Regularity” and “Start” and “Alarm mode” are weakly correlated to sleep quality.

In this section, I will also be presenting the performance of the 2 machine learning methods, and re-assess their suitability. Then, I will be moving on to comparing how the methods perform in the concatenated data frame and the data frame 2.



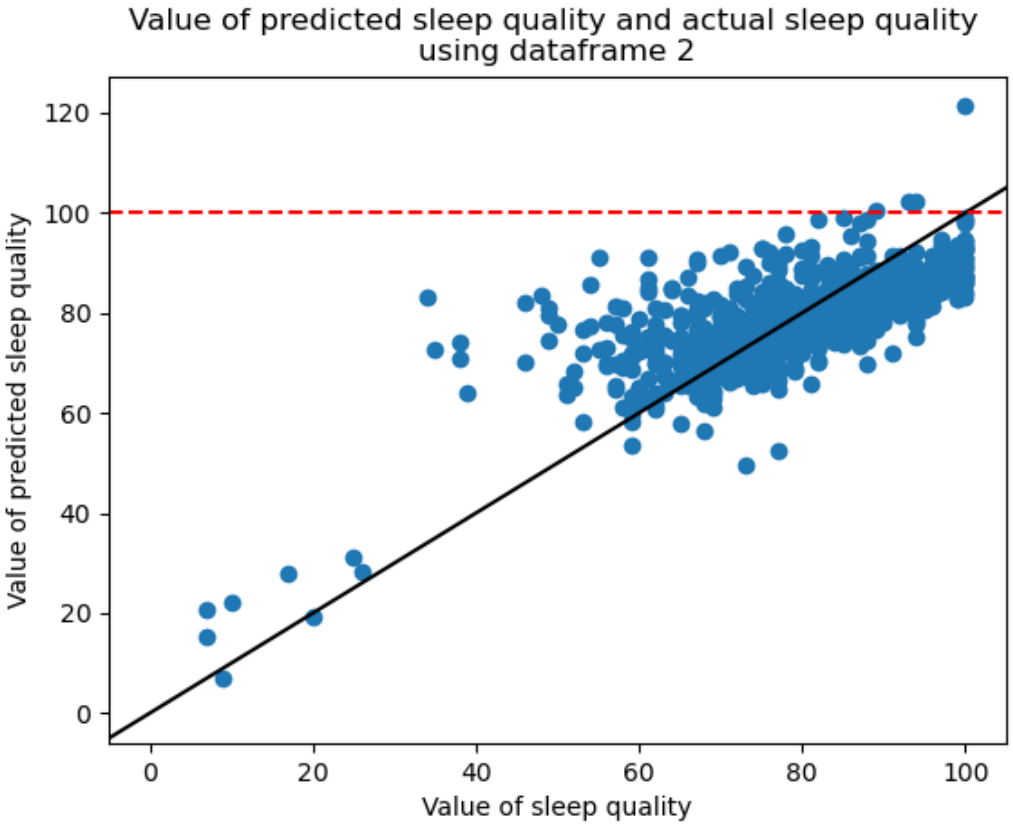
The plot above is the scatter plot of the value of the predicted sleep quality against the value of sleep quality. The black diagonal where the independent variables are equal to the dependent variables was added to aid the performance assessment. The mean squared error in this case is 106.37 (2 decimal places). This means the real values and predicted values are on average a bit more than 10% apart. Hence, the linear regression model can be used in predicting sleep quality with a fairly low error. However, we can also see that some of the predicted data are larger than 100, with the highest being 120. This is not a good sign, as sleep quality is a percentage and thus, cannot exceed 100 in value. This overshooting is likely due to the dataset not having sufficient lower-end data, while having a lot of upper-end data.

The accuracy score for the logistic regression model is 0.67 (2 decimal places), while the precision score is 0.65 (2 decimal places). This means the mode; yield a fairly reliable result.

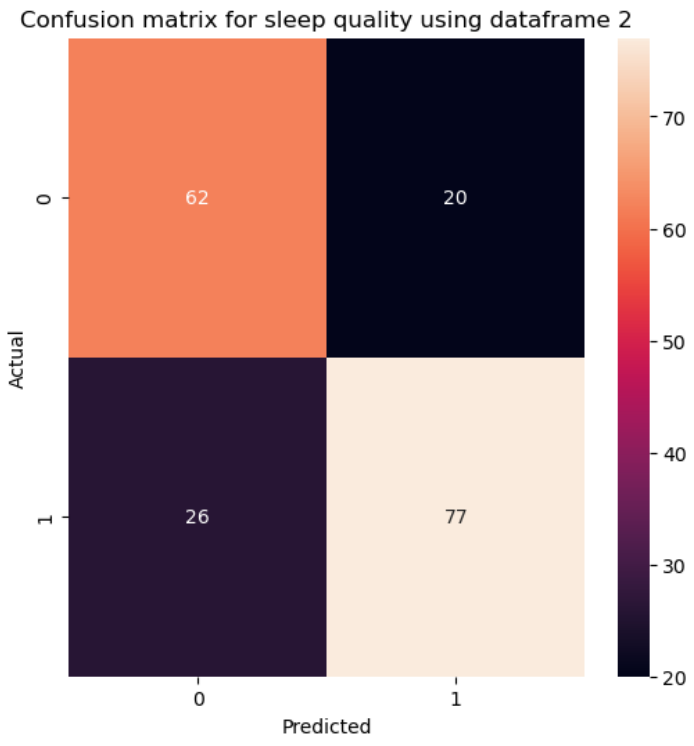


In our case of sleep analysis, there are no actual consequences for false positives and false negatives; therefore, this model's performance is rather satisfactory.

Let us look at how these models perform with the data frame 2. In the scatter plot below, we can see that the points are closer to the line where the independent variables are equal to the dependent variables. Moreover, the number of points that have been projected over the 100 threshold is roughly half of that when the linear regression model is implemented for the concatenated data frame. The mean squared error in this case is 79.38 (2 decimal places), meaning that there is around less than 9% difference between the actual values and the predicted values. From this, it is safe to conclude that linear regression works better with data frame 2 than the concatenated data frame.



When implemented on data frame 2, the logistic regression model yields the accuracy score of 0.75 (2 decimal places) and the precision score of 0.79 (2 decimal places). The confusion matrix yielded also reflects a better performance.



6. Conclusion & Discussion

From our result, we can conclude that it is possible to utilise the machine learning methods linear regression and logistic regression in predicting a person's general quality of sleep using their daily personal information. On one hand, this opens the possibility for the development of more services in industries such as the wellness industry. On another hand, this means that it is necessary to be wary of how one's personal information is handled, as it can reveal a lot about one's living habits and potentially, health conditions.

Choosing which model to use would mostly depend on the prediction needs. If there is a need to predict a continuous numerical value, linear regression should be chosen. If it is only necessary to predict whether or not the sleep quality is good, logistic regression should be employed.

We can also see how both models perform better when implemented on data frame 2. The concatenated data frame has significantly more data points (1808 data points) compared to data frame 2 (921 data points). However, data frame 2 allows for the use of more features in training the models than the concatenated data frame. In this case, it is possible to say that having more features allows for better performing models than having more data points.

The dataset used in this paper posed some limitations to data analysis. Firstly, this is not a big dataset. A bigger dataset would allow for a better training of data, and thus, a better result. Secondly, the data frames in this dataset do not have the same columns: data frame 2 has a more complete list of columns compared to that of data frame 1. Thirdly, this dataset is built using the data collected from one single subject; thus, it cannot be used as a representative dataset for the general population. Only consisting of a subject's data is also likely to be the reason why there are so few data points where the value of sleep quality falls into the lower-end. Moreover, every individual's health needs as well as their ability to enter deep sleep can differ greatly, making this dataset even less reliable.

For future analysis, it would have been better to have the data from the general population, instead of just one individual. This way, we will be able to catch a glimpse into how behaviours, bodily reactions, or the environment affect a human's sleep quality. Another important point for future analysis would be to have a more complete set of data, as the missing values in this dataset have taken away the opportunity to look at how some factors, such as weather type and air pressure, are correlated with sleep quality.

7. References

1. Alnawwar, M. A., Alraddadi, M. I., Algethmi, R. A., Salem, G. A., Salem, M. A., & Alharbi, A. A.. (2023). The Effect of Physical Activity on Sleep Quality and Sleep Disorder: A Systematic Review. Cureus. https://doi.org/10.7759/cureus.43595
2. Alotaibi, A. D., Alosaimi, F. M., Alajlan, A. A., & Bin Abdulrahman, K. A. (2020). The relationship between sleep quality, stress, and academic performance among medical students. Journal of family & community medicine, 27(1), 23–28. https://doi.org/10.4103/jfcm.JFCM_132_19
3. Del Brutto, O. H., Mera, R. M., Zambrano, M., & Castillo, P. R. (2016). Caffeine intake has no effect on sleep quality in community dwellers living in a rural Ecuadorian village (The Atahualpa Project). Sleep science (Sao Paulo, Brazil), 9(1), 35–39. https://doi.org/10.1016/j.slsci.2015.12.003
4. Diotte, D. (2022, April 25). Sleep data. Kaggle. https://www.kaggle.com/datasets/danagerous/sleep-data?select=sleepdata_2.csv
5. Jung, A. (2022). Machine learning the basics. Springer Nature.
6. Kang, J. H., & Chen, S. C. (2009). Effects of an irregular bedtime schedule on sleep quality, daytime sleepiness, and fatigue among university students in Taiwan. BMC public health, 9, 248. https://doi.org/10.1186/1471-2458-9-248
7. Sleep Cycle. (2023). Privacy notice https://www.sleepcycle.com/privacy-policy-2021/#:~:text=When%20using%20the%20Sleep%20Cycle,as%20snoring%20or%20other%20noises
8. Snel, J., & Lorist, M. M. (2011). Effects of caffeine on sleep and cognition. Progress in brain research, 190, 105–117. https://doi.org/10.1016/B978-0-444-53817-8.00006-2
9. Stull, R. (2019). ATSC 113 weather for sailing, flying & snow sports. UBC ATSC 113 - Standard Atmosphere-Pressure. https://www.eoas.ubc.ca/courses/atsc113/flying/met_concepts/02-met_concepts/02a-std_atmos-P/index.html
10. Sullivan Bisson, A. N., Robinson, S. A., & Lachman, M. E. (2019). Walk to a better night of sleep: testing the relationship between physical activity and sleep. Sleep health, 5(5), 487–494. https://doi.org/10.1016/j.sleh.2019.06.003
11. USDA. (2004). Fooddata Central Search Results. FoodData Central. https://fdc.nal.usda.gov/fdc-app.html#/food-details/171890/nutrients
12. USDA. (2005). Fooddata Central Search Results. FoodData Central. https://fdc.nal.usda.gov/fdc-app.html#/food-details/174873/nutrients