A Prediction Study on Sleep Efficiency

Group: L2D - 6
Members: Chunyu Zhang, Ece Celik, Rashi Selarka, Shiyu Jiang

INTRODUCTION

In recent years, research has shown that nearly 60% of all university students suffer from poor sleep quality and 7.7% of them meet the criteria for insomnia (Schlarb, A. A., Friedrich, A., & Claßen, M, 2017). Sleeping problems in students have direct consequences such as depression, reduced life satisfaction, irritability and poor academic performance . Past research done in the field also provides evidence for the positive correlation between academic failure and poor sleep quality patterns (Centers for Disease Control and Prevention, 2022).

Considering these negative impacts of poor sleep quality we decided to explore a Kaggle dataset that contains information on a group of test subjects and their sleep patterns to see which variables have a relationship with poor sleep quality. The dataset collected information on 452 subjects' age, gender, waking and bedtimes, proportions of time spent in the various sleep stages, and habits (smoking, alcohol consumption, etc.).

The dataset can be found at the following link:
https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency

By studying past research, we found that age, alcohol intake, smoking and frequency of exercise are all associated with sleep quality. So, we believe that sleep quality is related to these variables. To predict the sleep efficiency of college students, we decided to explore their relationship with sleep efficiency. In our study, smoking levels were not used because college students had fewer smokers and, in comparison, higher alcohol intake.

When assessing sleep quality, we noticed that age, alcohol, and exercise are often linked to the Rapid Eye Movement (REM) (WebMD Editorial Contributors, 2022) stage of the sleep cycle, and so we wished to explore whether there is a certain relationship between age, exercise frequency, alcohol consumption and percentage of REM sleep. We have thus decided to examine the predictive question: ***How do age, exercise frequency and alcohol consumption predict the percentage of REM sleep?***

To answer our research question, we will be using the following three explanatory variables from our dataset:

- the age of the test subject measured in years ('Age') (a continuous variable)
- the number of times the test subject exercises each week ('Exercise.frequency') (a discrete variable)
- the amount of alcohol consumed within 24 hours before bedtime (measured in fl oz) ('Alcohol.consumption') (a continuous variable)

As our response variable, we will use:

- the percentage of total sleep time spent in the REM stage ('REM.sleep.percentage') (a continuous variable).

Considering these variables, we wanted to predict the REM sleep percentage for an *average* college student, and so we decided to predict the REM performance for a 20 year old (the average of 18-22 year olds) exercising 3 times a week (Billitz, 2023) and having 2.4 fl oz of alcohol (equivalent to 4 drinks and this is the average number of drinks for female college student) (Sobering Up, 2021).

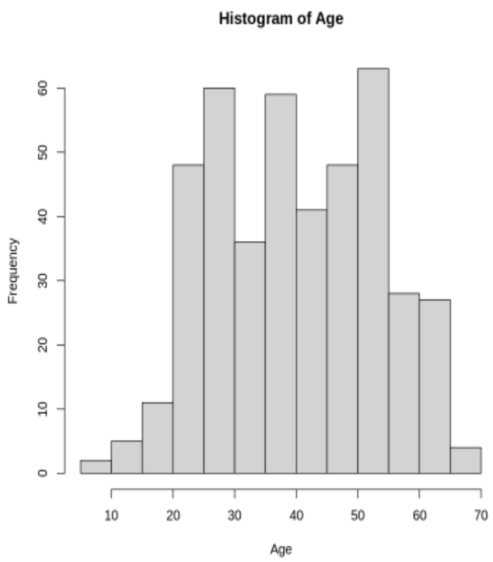
The reason we decided to select these variables in particular for exploration is because they seemed to be the most relevant behaviours and factors to the average college student's sleep quality, and we then decided to research and study those assumptions. It is a widely regarded phenomenon that with age, one's sleep quality and the REM proportion of it reduces (Ohayon et al., 2004)(Pótári et al., 2017)(Van Cauter, 2000), and why we thought it would be interesting to explore the variable of age. It would also serve as somewhat of a point of comparison or control in our model since we feel most certain about it having an inverse relationship with REM percentage. Since the maximum REM percentage in the dataset is 30, we hypothesize a 20 year old's best REM sleep percentage would be around 25. With the other variables, exercise and alcohol, we've found research and arguments supporting either side. While there is some debate on the specific effects general exercise can have on sleep (Kripa & Jackson, n.d.), some studies have found that exercising at night or within 4 hours of bedtime can lead to reduced REM activity (Breus, 2022)(Falk, 2022)(Frimpong et al., 2021). Since a lot of college students tend to exercise in the evenings (due to classes during the day), we are assuming for our prediction that the person was exercising within 4 hours of sleeping. Similarly for alcohol, there is much research to support that increased amounts of alcohol consumption lead to spending less time in the REM stage during the sleep cycle (Pacheco, 2023)(Thakkar, Sharma & Sahota, 2015)(Colrain, Nicholas & Baker, 2014). *Therefore, we hypothesize that a 20 year old exercising thrice a week (within 4 hours of sleeping) and having 2.4 fl oz of alcohol within 24 hours of sleeping would reduce their REM sleep percentage to around 21.*

To make our prediction, we fit a multiple linear regression model by using the age, alcohol consumption, and exercise frequency as the explanatory variables and the REM sleep percentage as the response variable.

We performed linear regression to fit an additive multilinear model. To evaluate the model's performance, we examined the summary of the model, which included information such as coefficient estimates, standard errors, t-values, p-values, and R-squared values. Additionally, we also fit an interactive multilinear model to compare the two models' predictions. To compare the performance of the different models, we used residual plots and QQ plots to evaluate the model's fit and identify potential issues such as outliers or heteroscedasticity. We kept in mind that the additive model would work best when the independent variables exist and act independently, while the interactive model would work best when there are complex interactions between the explanatory variables.

ANALYSIS

The three histograms below show how the three explanatory variables are distributed. The first histogram illustrates the majority of the data is between age 20 to 55. The second histogram shows the most recurrent exercise frequencies are 0, 1, & 3 in a week. The third one shows over half of the people in the dataset have no alcohol consumption, and the remainder are distributed almost evenly through the rest of the levels.



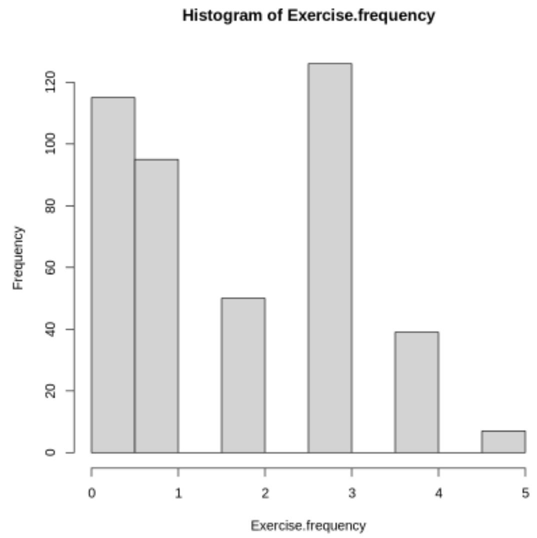


Figure 1: Distribution of Age

Figure 2: Distribution of Exercise Frequency

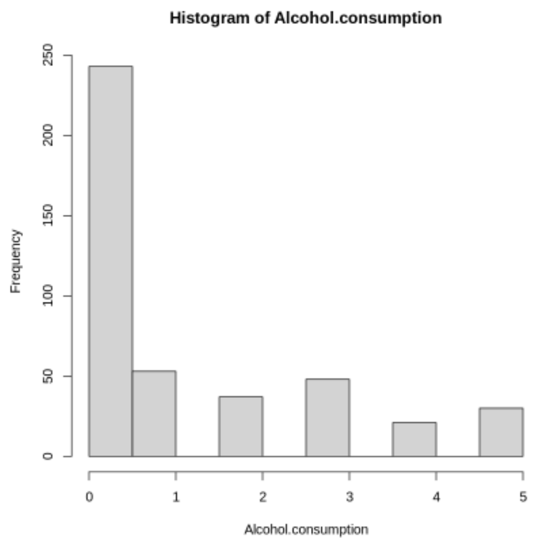


Figure 3: Distribution of Alcohol Consumption

We started our analysis by building two linear models. First, we created an additive model and used the R function `summary ()` to view our model summary. The model summary for our additive model can be seen below.

```
Call:
lm(formula = REM.sleep.percentage ~ Age + Alcohol.consumption +
    Exercise.frequency, data = sleep_training)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1437 -2.7490 -0.2389  2.2479  7.1212 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         22.6303654  0.6749743  33.528   <2e-16 ***
Age                  0.0003348  0.0149198   0.022    0.982    
Alcohol.consumption -0.1361768  0.1202328  -1.133    0.258    
Exercise.frequency   0.1240768  0.1413342   0.878    0.381    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.444 on 297 degrees of freedom
Multiple R-squared:  0.006828,  Adjusted R-squared:  -0.003204 
F-statistic: 0.6807 on 3 and 297 DF,  p-value: 0.5645
```

Figure 4: Summary of additive model

We observed that in our additive model the adjusted R square is -0.003204. We see that the value is a negative number, which suggests that the additive model we used here fits the data very poorly.

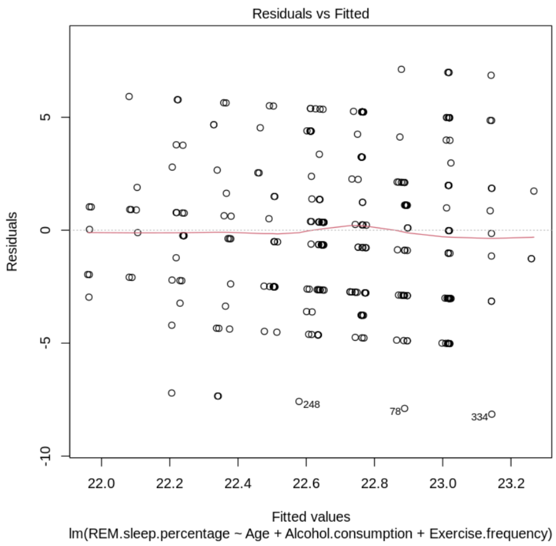
So, we also built an interactive model to see if that can increase the R-squared value that we are getting. Below we have the summary of our interactive model.

```
Call:
lm(formula = REM.sleep.percentage ~ Age * Alcohol.consumption *
    Exercise.frequency, data = sleep_training)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0874 -2.7421 -0.0277  2.1230  7.1229 

Coefficients:
                                          Estimate Std. Error t value   Pr(>|t|)    
(Intercept)                              23.048992   1.063074  21.681     <2e-16 ***
Age                                      -0.011733   0.025685  -0.457      0.648    
Alcohol.consumption                       0.200660   0.641871   0.313      0.755    
Exercise.frequency                        0.289961   0.633153   0.458      0.647    
Age:Alcohol.consumption                  -0.006456   0.014684  -0.440      0.661    
Age:Exercise.frequency                   -0.003470   0.014566  -0.238      0.812    
Alcohol.consumption:Exercise.frequency   -0.457451   0.304804  -1.501      0.134    
Age:Alcohol.consumption:Exercise.frequency  0.010703   0.006989   1.532      0.127    

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.437 on 293 degrees of freedom
Multiple R-squared:  0.0243,    Adjusted R-squared:  0.0009881 
```

Figure 5: Summary of interactive model

We observed that in our interactive model the adjusted R-squared is 0.0009881. While an improvement from the additive model, it is still a very small value.

To have a better understanding of our models, we created residual plots and QQ-plots for both our additive and interactive model. The residual plot and qq-plot for our additive model can be seen below.



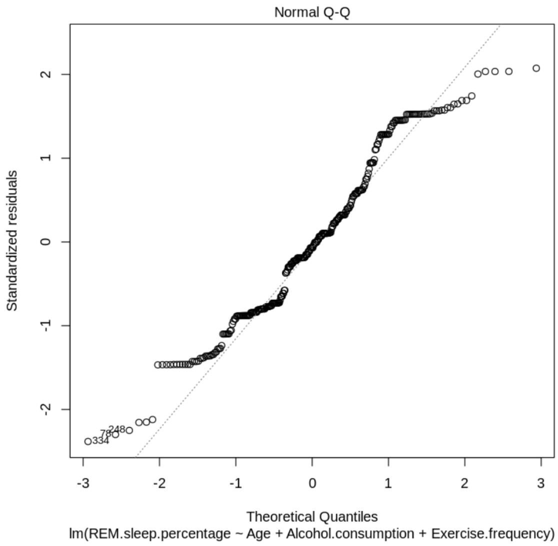
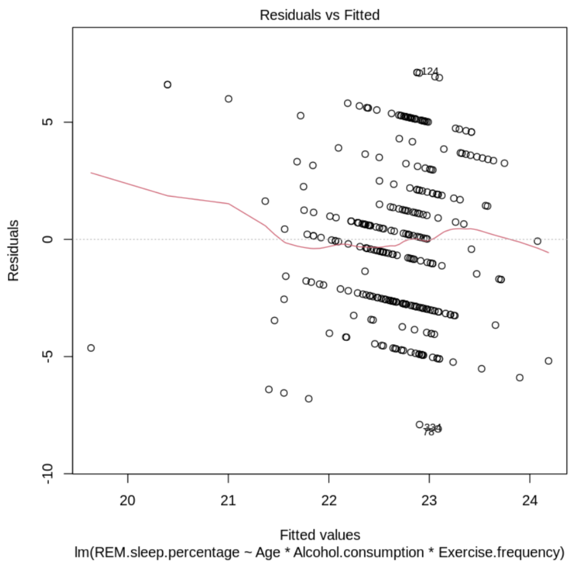


Figure 6: Residual plot for additive model

Figure 7: QQ-plot for additive model

Here we observed that there is some sort of randomness in the residual plot which can indicate that our model is a good fit. However, since we observed a negative R-squared value for our additive model, we also checked our qq-plot for the additive model. In the qq-plot of residuals we see that there are significant deviations from the line and especially in the tails. This suggests that our residuals do not follow a normal distribution. Considering that we assume normality in linear regression models, this violation of normality is another sign that the additive model is not a good fit for our data.

We also created a residual plot and qq-plot for our interactive model as it can be seen below.



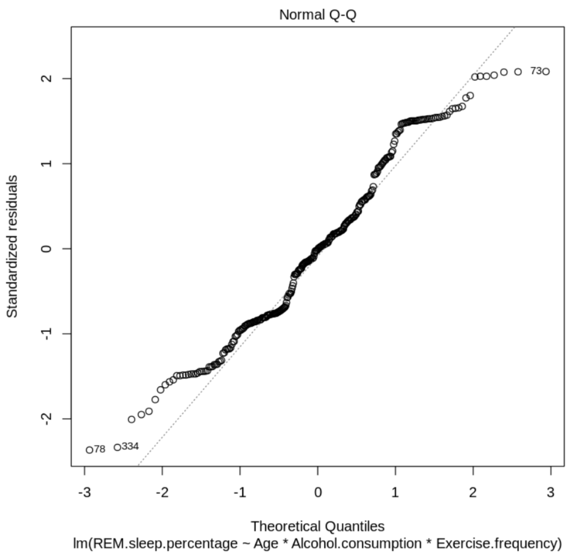


Figure 8: Residual plot for interactive model

Figure 9: QQ-plot for interactive model

In the residual plot for the interactive model we observed that there is not a random pattern that we would like to see in good linear models. This result supports our low R-squared value and indicates that our additive model is not a good fit. In addition, when we generated our qq-plot for the interactive model we again saw that there are significant deviations from the line and especially in the tails. This suggests that our residuals do not follow a normal distribution and this violates the normality assumption in linear models. Thus considering all of our results, we saw that our interactive model was also not a good fit for our data.

While both models didn't seem apt, we wanted to see if one might still be better than the other. By computing the RMSE (which measures the average difference between the predicted and actual values to provide an estimation of how accurate the model is), we can compare which of the two models fit the data better. The RMSE found for the additive model was 3.517 and the RMSE for the interactive model was 3.551. Moreover, we also evaluate the two models by the AIC (Akaike Information Criterion). The AIC of the additive model and the interactive models were 1598.051 and 1600.017 respectively.

Upon setting up the models and computing any descriptive statistics that could help us assess the accuracy of the models, we moved onto the main investigation of our project, which was to predict the REM sleep percentage for a 20 year old exercising thrice a week and having 2.4 fl oz of alcohol (4 drinks) within 24 hours of sleeping. We used the `predict` function in R to do this.

The **additive model predicted** the REM sleep percentage to be **22.236** and the **interactive model predicted** it to be **20.947.**

DISCUSSION

Upon fitting the additive and interactive models to our data, we found that there wasn't actually that large a difference between the two models' predictive power. We computed the AIC for both models, and found them again to be fairly similar to each other for every randomly chosen training data. Since the AIC usually penalizes models with more parameters, we would expect the interaction model to have a larger AIC, but the fact that both models' AIC values are almost the same, it shows that there isn't a difference in the models' predictive abilities.

The RMSEs for both models were definitely on the lower end (and extremely similar to each other (3.52 for additive vs. 3.55 for interactive), suggesting the models are fairly accurate at predicting REM sleep percentage. In fact, we initially hypothesized that for a 20 year old exercising thrice a week (within 4 hours of sleeping) and having 2.4 fl oz of alcohol within 24 hours of sleeping, the REM sleep percentage would be around 21, and the predicted values computed from the additive and interactive models were 22.236 and 20.947 respectively, making our hypothesis nearly right in the middle of the two.

It is certainly clear to see that our hypothesized value is (marginally) closer to the interactive model than to the additive, and it only makes sense that it is. Interaction between our chosen explanatory variables being present is something that is easily inferred, but also has much research to support it. There is much research supporting the fact that alcohol consumption and exercise are heavily related, because increased alcohol consumption can lead to dehydration and fatiguing, and thus lower levels of exercise (El-Sayed, Ali & El-Sayed Ali, 2005); the flipped effects have also been investigated, wherein people who exercise more were found to have higher tolerance and consume larger amounts of alcohol (Conroy et al., 2015) (Leichliter et al., 1998). We also know that with age, exercise frequency usually goes down (Woo et al., 2006), and that age and alcohol consumption have been found to have varying relationships demographically (Eigenbrodt et al., 2001). Therefore, this would explain why including interaction terms would provide a more accurate predictive model for this data and this set of chosen variables.

Speaking of the chosen variables however, adjusted R squared for both models are rather low, suggesting these variables in a linear regression may not necessarily be the best fit to the data. It would suggest that perhaps if we are to fit a linear model, we would have to go about selecting different variables for our model, or if these variables are the ones to be used, it would be more fitting in a non-linear regression. We attempted to use the model selection algorithm via `regsubsets` to explore this and got the following response:

```
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         Age Alcohol.consumption Exercise.frequency Age:Alcohol.consumption
1  ( 1 ) " " " "                 " "                " "
2  ( 1 ) " " " "                 " "                " "
3  ( 1 ) " " " "                 " "                "*"
4  ( 1 ) " " "*"                 " "                "*"
5  ( 1 ) " " "*"                 "*"                "*"
6  ( 1 ) " " "*"                 "*"                "*"
7  ( 1 ) "*" "*"                 "*"                "*"
         Age:Exercise.frequency Alcohol.consumption:Exercise.frequency
1  ( 1 ) "*"                    " "
2  ( 1 ) " "                    "*"
3  ( 1 ) " "                    "*"
4  ( 1 ) " "                    "*"
5  ( 1 ) " "                    "*"
6  ( 1 ) "*"                    "*"
7  ( 1 ) "*"                    "*"
         Age:Alcohol.consumption:Exercise.frequency
1  ( 1 ) " "
2  ( 1 ) "*"
3  ( 1 ) "*"
4  ( 1 ) "*"
5  ( 1 ) "*"
6  ( 1 ) "*"
7  ( 1 ) "*"
```

Figure 8: Best models for various number of explanatory variables

This algorithm suggests that for a linear model with 3 explanatory variables, the 3 that should be chosen are the interaction term between age and alcohol consumption, exercise and alcohol consumption, and all 3 of them. Since it wouldn't make sense to have a model with interaction terms and not the main terms, it leads us to believe we are better off fitting a non-linear model for these 3 explanatory variables.

CONCLUSİON

In this project before we started building our models we did some research to decide which explanatory variables to choose for our model. In our research we saw that there is a relationship between age, alcohol consumption and exercise frequency and sleep efficiency, and so we wanted to explore the question of "***How do age, exercise frequency and alcohol consumption predict the percentage of REM sleep?"*** .

However when we built our additive model we observed a negative R-squared (-0.003204), which indicated a very poor fit. Then we created residual and qq-plots to further explore our model and we observed a qq-plot that indicates a violation in the residual normality assumption. These results for our additive models led us to conclude that our additive model was not a good fit at all.Then we created our interactive model but also observed a very low R-squared (0.0009881). When we created our residual and qq-plots for the interactive model we saw similar results to our additive model and concluded that our interactive model was not a good fit either. Therefore we can say that there was not a significant difference in the predictive abilities of our models. We also calculated RMSE and AIC values. Even though our additive model had a smaller RMSE (3.517 vs. 3.551), considering the other results we had for our additive model, we can not say that this is an indication for a better model. For our AIC values, we observed that our additive model had 1598.051 and our interactive model 1600.017. These were very high values, which lend support to the idea that these models are not ideal. So, we realized in the beginning that we did not have very potent models to answer our chosen research question.

Nonetheless, though our models did not fit the data as well as we would've wanted them to, we still wanted to see how they would answer the main question of our project, and if we they could predict the REM sleep percentage for a 20 year old exercising thrice a week and having 2.4 fl oz of alcohol (4 drinks) within 24 hours of sleeping as we hypothesized. The additive model predicted the REM sleep percentage to be 22.236 and the interactive model predicted it to be 20.947. These values are close to our hypothesized value that we came up with (REM Percentage = 21) only through educated guessing based on research. We observe that our hypothesized value is closer to the predicted value from our interactive model. This made sense to us considering that regardless of the R-squared for the interactive model being very low, it was marginally better than our additive model. However this wasn't enough to conclude that we have a good interactive model that could fit the data well; this closeness in the hypothesis and prediction is most likely due to happenstance.

Therefore, we started to think about what we could do to improve these models as a response to our research question in the future. Considering the results we obtained from this project, a potential reason why we did not get good models could be owing to the number of variables we chose. Future researchers can try to explore if adding more explanatory variables in the model can increase the model accuracy. The variables from our dataset that we did not include from our dataset such as "Sleep.duration", "Deep.sleep.percentage", "Light.sleep.percentage", "Caffeine.consumption" can be included in new models. Since in our analysis we saw that linear regressions do not fit the data well, projects that want to continue only working on the three variables that we used in this investigation should consider using non-linear models to fit the data.

An interesting future research topic can be “even when we keep the same number of explanatory variables (3 variables) how would the model accuracy change with different explanatory variables”. As we mentioned in our discussion above, Figure 8 suggests for 3 explanatory variables that we should be considered to have the best linear model are:

-Age:Alcohol.consumption (age-alcohol consumption interaction)
-Alcohol.consumption:Exercise.frequency(alcohol consumption - exercise frequency interaction)
-Age:Alcohol.consumption:Exercise.frequency (age - alcohol consumption - exercise frequency interaction)

So, in the future a project can be done where somehow only these three variables, or a group of any 3 others from this dataset, are considered as explanatory variables in a linear model and the model accuracy can be compared to the models’ accuracies that we created in this project.

---end of report---

References

Alcohol and sleep. Drinkaware. (n.d.). Retrieved March 16, 2023, from https://www.drinkaware.co.uk/facts/health-effects-of-alcohol/effects-on-the-body/alcohol-and-sleep

Billitz, J. (2023, January 29). *17 college student exercise statistics (rates & factors)*. NOOB GAINS. Retrieved March 17, 2023, from https://www.noobgains.com/exercise-statistics-college-students/

Breus, D. M. (2022, December 13). *Exercise and sleep*. The Sleep Doctor. Retrieved March 16, 2023, from https://thesleepdoctor.com/exercise/

Centers for Disease Control and Prevention. (2022, September 15). *Psychosocial correlates of insomnia among college students*. Centers for Disease Control and Prevention. Retrieved March 17, 2023, from https://www.cdc.gov/pcd/issues/2022/22_0060.htm#:~:text=However%2C%20at%20least%2060%25%20of,poor%20sleep%20quality%20(3)

Colrain, I. M., Nicholas, C. L., & Baker, F. C. (2014). Alcohol and the sleeping brain. *Handbook of clinical neurology*, *125*, 415–431. https://doi.org/10.1016/B978-0-444-62619-6.00024-0

Conroy, D. E., Ram, N., Pincus, A. L., Coffman, D. L., Lorek, A. E., Rebar, A. L., & Roche, M. J. (2015). Daily physical activity and alcohol use across the adult lifespan. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, *34*(6), 653–660. https://doi.org/10.1037/hea0000157

Eigenbrodt, M. L., Mosley, T. H., Hutchinson, R. G., Watson, R. L., Chambless, L. E., & Szklo, M. (2001). Alcohol consumption with age: A cross-sectional and longitudinal study of The atherosclerosis risk in communities (ARIC) study, 1987-1995. *American Journal of Epidemiology*, *153*(11), 1102–1111. https://doi.org/10.1093/aje/153.11.1102

El-Sayed, M. S., Ali, N., & El-Sayed Ali, Z. (2005). Interaction between alcohol and exercise: physiological and haematological implications. *Sports medicine (Auckland, N.Z.)*, *35*(3), 257–269. https://doi.org/10.2165/00007256-200535030-00005

Falk, M. (2022, November 29). *How working out before bed can affect your sleep*. Shape. Retrieved March 16, 2023, from https://www.shape.com/exercise-before-bed-sleep-6833389#:~:text=Plus%2C%20high%2Dintensity%20exercise%20ending,to%20a%202021%20meta%2Danalysis

Frimpong, E., Mograss, M., Zvionow, T., & Dang-Vu, T. T. (2021). The effects of evening high-intensity exercise on sleep in healthy adults: A systematic review and meta-analysis. *Sleep medicine reviews*, *60*, 101535. https://doi.org/10.1016/j.smrv.2021.101535

Kripa, S., & Jackson, K. (n.d.). *Effects of exercise on sleep*. Physiopedia. Retrieved March 16, 2023, from https://www.physio-pedia.com/Effects_of_Exercise_on_Sleep#:~:text=Exercise%20Duration,-It%20is%20necessary&text=A%20meta%2Danalytical%20study%20shows,than%201%20hour%20a%20day

Leichliter, J. S., Meilman, P. W., Presley, C. A., & Cashin, J. R. (1998). Alcohol use and related consequences among students with varying levels of involvement in college athletics. *Journal of American College Health*, *46*(6), 257–262. https://doi.org/10.1080/07448489809596001

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. V. (2004). Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*, *27*(7), 1255–1273. https://doi.org/10.1093/sleep/27.7.1255

Pacheco, D. (2023, February 8). *Alcohol and sleep*. Sleep Foundation. Retrieved March 16, 2023, from https://www.sleepfoundation.org/nutrition/alcohol-and-sleep

Pótári, A., Ujma, P. P., Konrad, B. N., Genzel, L., Simor, P., Körmendi, J., Gombos, F., Steiger, A., Dresler, M., & Bódizs, R. (2017). Age-related changes in sleep EEG are attenuated in highly intelligent individuals. *NeuroImage*, *146*, 554–560. https://doi.org/10.1016/j.neuroimage.2016.09.039

Schlarb, A. A., Friedrich, A., & Claßen, M. (2017, July 26). *Sleep problems in university students - an intervention*. Neuropsychiatric disease and treatment. Retrieved March 17, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5536318/

Sobering Up. (2021, August 26). *Infographic: How much do college students drink?* Sobering Up. Retrieved March 17, 2023, from https://www.scramsystems.com/blog/2014/02/infographic-much-college-students-drink/

Thakkar, M. M., Sharma, R., & Sahota, P. (2015). Alcohol disrupts sleep homeostasis. *Alcohol*, *49*(4), 299–310. https://doi.org/10.1016/j.alcohol.2014.07.019

U.S. Department of Health and Human Services. (n.d.). *What is a standard drink?* National Institute on Alcohol Abuse and Alcoholism. Retrieved March 16, 2023, from https://www.niaaa.nih.gov/alcohols-effects-health/overview-alcohol-consumption/what-standard-drink#:~:text=Each%20beverage%20portrayed%20above%20represents,14%20grams%20of%20pure%20alcohol

Van Cauter, E. (2000). Age-related changes in slow wave sleep and REM sleep and relationship with growth hormone and cortisol levels in healthy men. *JAMA*, *284*(7), 861. https://doi.org/10.1001/jama.284.7.861

WebMD Editorial Contributors. (2022). *Stages of sleep: Rem and Non-REM Sleep cycles*. WebMD. Retrieved March 17, 2023, from https://www.webmd.com/sleep-disorders/sleep-101

Woo, J. S., Derleth, C., Stratton, J. R., & Levy, W. C. (2006). The influence of age, gender, and training on Exercise Efficiency. *Journal of the American College of Cardiology*, *47*(5), 1049–1057. https://doi.org/10.1016/j.jacc.2005.09.066