Exploring the Relationship Between Sleep Duration and College GPA

Tasnim Rida

October 13, 2023

Introduction

Many college students experience issues with their sleep schedule. Balancing rigorous academics, social life, and personal interests can be difficult, leaving little time for sleep. Especially as a college student, academic performance is often prioritized over health and well-being. As such, many students stay up late to complete assignments and wake up early for morning classes. Given that less sleep is linked with to various physical and psychological health deficits, it is theorized that sleep time could have a relationship with academic performance. The goal of this analysis is to find whether or not there is an association between sleep time and academic performance, specifically GPA. This analysis may be useful in deciding whether or not early morning classes should be eliminated, depending on whether the analysis reveals that less sleep is correlated with lower GPA.

To collect the data, researchers had students from Carnegie Mellon University and two other universities wear sleep trackers to track average sleep times per night for one month during the spring semester. The average minutes of sleep per night measures sleep time. Researchers then collected student's semester grades as well as their cumulative GPA from previous semesters. In this context, GPA is a measure of academic performance. The analysis found that sleep time and GPA have a positive relationship, providing evidence that less sleep is associated with a lower GPA. Thus, student schedules should consider this to improve student work-life balance and academic performance.

Exploratory Data Analysis & Data Summary

The data used to answer the research questions comes from a study conducted at Carnegie Mellon University and two other universities. The dataset contains data on 634 students in total.

The variables of interest in this study are as follows:

- `TotalSleepTime`: the average time the student slept each night (in minutes)
- `term_gpa`: the student's semester GPA (out of 4.0)
- `cum_gpa`: the student's cumulative GPA (out of 4.0) for previous semesters

`TotalSleepTime` variable is a continuous quantitative variable while `term_gpa` and `cum_gpa` are discrete quantitative variables. To further explore the distributions of these three variables, exploratory data analysis was conducted.

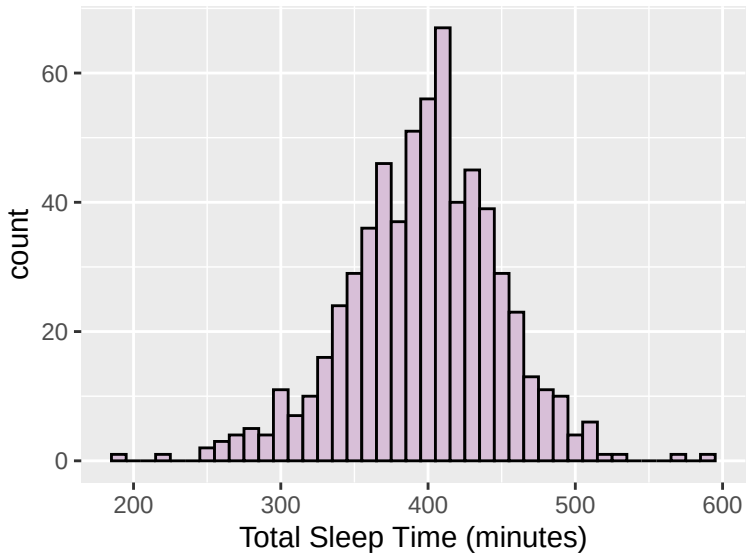

Figure 1: Marginal Distribution of Total Sleep Time.

Figure 1 displays the marginal distribution of the `TotalSleepTime` variable. The marginal distribution of the `TotalSleepTime` seems to be approximately normal, symmetric, and unimodal. The center of the distribution is around 400 minutes. There does not appear to be any apparent outliers.

Figure 2 displays the marginal distribution of the `term_gpa` variable. The marginal distribution of the `term_gpa` seems to be left skewed and unimodal. There does appear to be some outliers; in particular, the leftmost bar seems to be a possible outlier (representing a students with low GPAs).

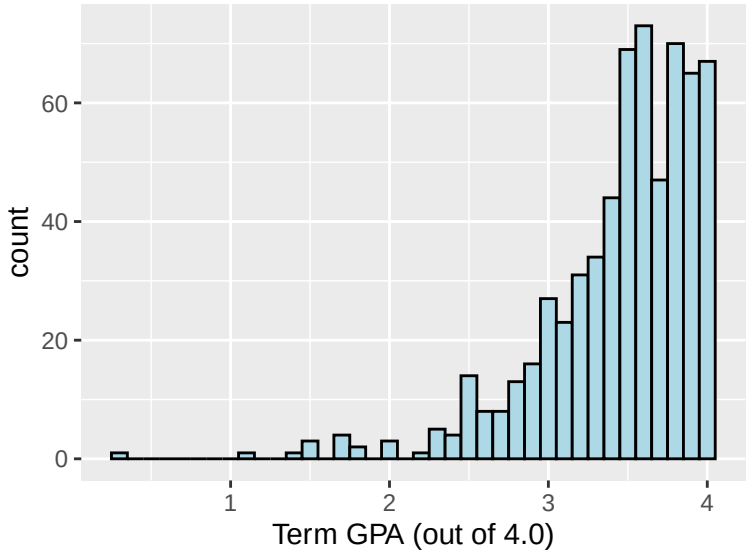

Figure 2: Marginal Distribution of term_gpa.

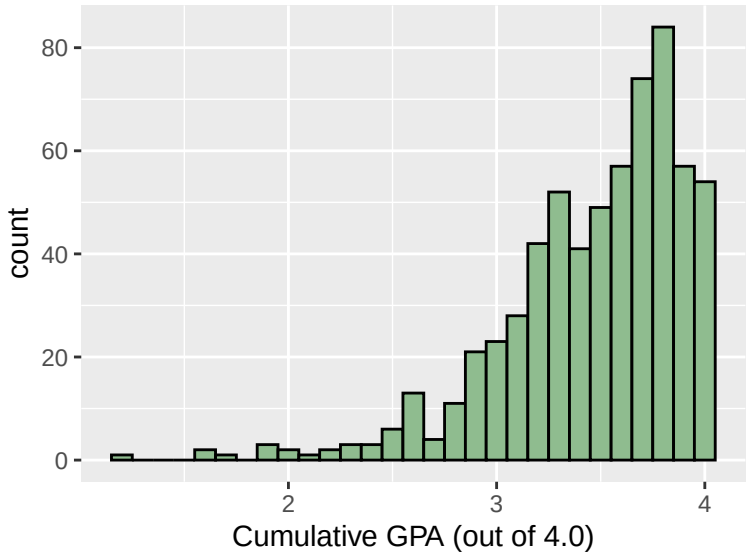

Figure 3: Marginal Distribution of cum_gpa.

Figure 3 displays the marginal distribution of the `cum_gpa` variable. The marginal distribution of the `term_gpa` seems to be left skewed and unimodal. There does appear to be some outliers; in particular, the leftmost bar seems to be a possible outlier (representing a students with low GPAs).

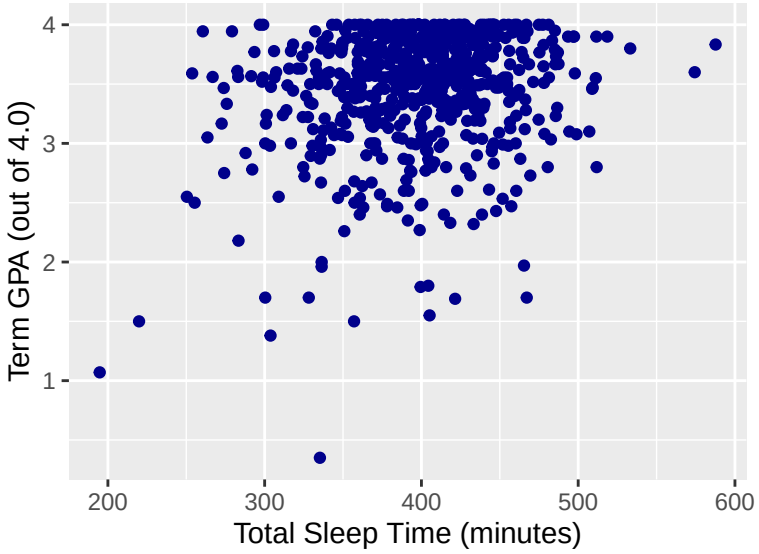

Figure 4: Total Sleep Time and Term GPA.

Figure 4 displays `TotalSleepTime` plotted against `term_gpa`. Upon initial inspection, there seems to be a strong, positive linear relationship between the two variables. It is important to note that there seem to be quite a few outliers, located to the top right of the plot (representing students who had high averages sleep each night and high term GPAs) as well as the bottom left of the plot (those who got lower averages of sleep and lower GPAs).

Figure 5 displays `TotalSleepTime` plotted against `cum_gpa`. Upon initial inspection, there seems to be a strong, positive linear relationship between the two variables. Although the plot looks similar to Figure 4, there seems to be more random scatter. Whereas in Figure 4 the points seem to be more concentrated around the top middle of the graph, Figure 5 has more scatter. There also seems to be a few outliers, located to the top right of the plot (representing students who had high averages sleep each night and high term GPAs) as well as the bottom middle of the plot (those who got an in-the-middle average of sleep and lower GPAs), and the middle left of the plot (those who got lower averages of sleep and GPAs that fall in the middle).

Based on these plots, the relationship I chose to model is between `TotalSleepTime` and `term_gpa`. Figure 4 displays a possibly linear relationship between these two variables.

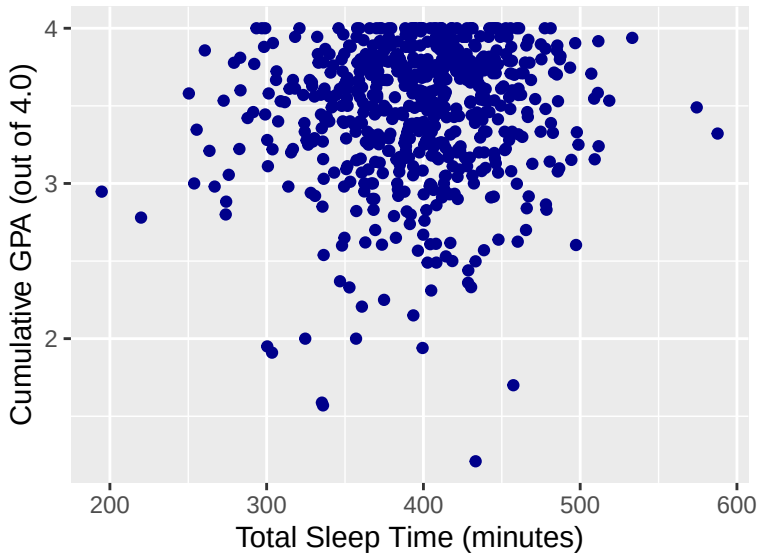

Figure 5: Total Sleep Time and Cumulative GPA.

Whilst there also seems to be a linear relationship between `TotalSleepTime` and `cum_gpa` (as indicated by Figure 5), the research question is focused on the association between sleep time and GPA. Since the variable `term_gpa` concerns the GPAs for the semester which students recorded their average sleep times, I chose to omit the `cum_gpa` variable.

Methods

Based on the exploratory data analysis, I chose to fit a simple linear regression model with the variables `TotalSleepTime` and `term_gpa`. A simple linear regression model in this context clearly models the association between sleep time and GPA.

However, the original `TotalSleepTime` data is in terms of minutes. Since the substantive question asks for the association between sleep time in hours and gpa, I chose to transform the `TotalSleepTime` variable by dividing by 60. This will result in values for `TotalSleepTime` in hours since there are 60 minutes in 1 hour.

Figure 6 displays the relationship between `TotalSleepTime` and `term_gpa` with the linear regression line plotted. This justifies the linearity assumption for simple linear regression, as the relationship visually looks plausibly linear. However, there is noticeable variation in the data that is not consistent with the linear regression line plotted, indicating potential problems in fitting a simple linear regression model.

The slope estimate for the regression line can help us answer the question of association

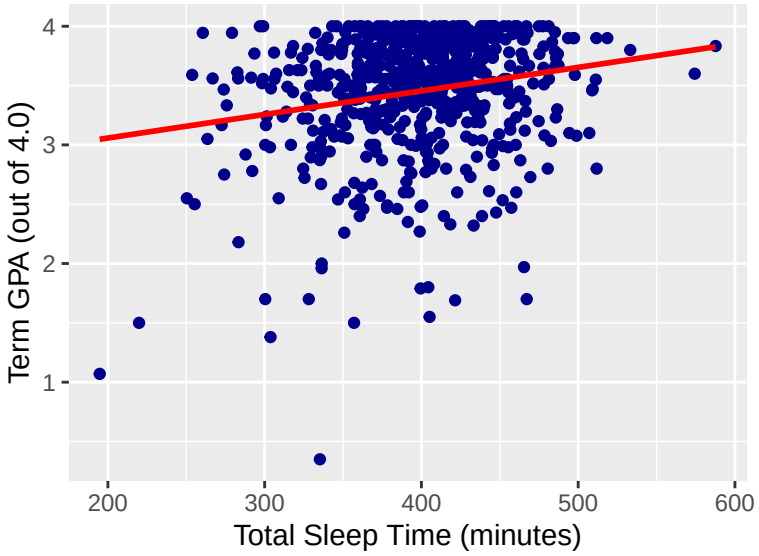

Figure 6: Total Sleep Time and Term GPA.

between sleep time and GPA; if the slope estimate is nonzero, there is evidence of a relationship between sleep time and GPA. A 95% confidence interval will also be constructed for the slope estimate such that the true slope value will fall between this interval.

Another question of interest is finding the average GPA effect with 2 hours less sleep, since students who have early 8AM classes are expected to sleep approximately 2 hours less. To calculate this average effect, the slope estimate will be multiplied by -2. Possible causation will also be evaluated.

Results

The linear regression model is as follows: term_gpa = beta1(sleep.time) + beta0, where beta1 is the slope estimate and beta0 is the intercept. The fitted linear regression model and its outputs are as follows:

```
##
## Call:
## lm(formula = term_gpa ~ sleep.time, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97642 -0.20934  0.09687  0.36328  0.76631
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6610     0.1536  17.327  < 2e-16 ***
## sleep.time    0.1191     0.0230   5.176 3.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4906 on 632 degrees of freedom
## Multiple R-squared:  0.04067,    Adjusted R-squared:  0.03916
## F-statistic:  26.8 on 1 and 632 DF,  p-value: 3.043e-07
```

In the case of linear regression, the null hypothesis is that beta1 = 0 whilst the alternative hypothesis is that beta1 does not equal zero. The results of the simple liner regression gives us an estimate of 0.1191 for beta1 and an estimate of 2.6610 for beta0. Since the p-value for the slope estimate 3.04x 10^-7 is less than 0.05, we reject the null hypothesis. Thus, there is statistical evidence that beta1, the slope coefficient, is not equal to 0, indicating that there is an association between sleep time and GPA.

```
##             Coefficient Lower_Bound Upper_Bound
## (Intercept) (Intercept)  2.35947169   2.9626284
## sleep.time   sleep.time  0.07390545   0.1642524
```

The table above displays the lower and upper bounds of a 95% confidence interval for the slope estimate. Thus, we are 95% confident that the true value of the slope falls within the interval [0.07390545, 0.1642524].

As stated before, based on the model, we can conclude that there is an association between sleep time (in hours) and GPA. The slope estimate indicates to us that: getting an additional hour of sleep is associated with having a GPA that is beta1 = 0.1191 units higher, on average (95% CI [0.07390545, 0.1642524]. Thus, there is evidence that students who sleep less indeed get lower GPAs because the relationship is positive, meaning that as sleep time increases, GPA increases.

The average GPA effect with 2 hours less sleep was calculated to be 0.1191 * -2 = -0.2382. This indicates that: getting an additional hour of sleep is associated with having a GPA that is beta1 = 0.2382 units lower, on average. It is also important to note that we cannot conclude that GPA change is caused by less sleep; linear regression does not model causation. This is because there could be various other factors that cause GPA such as being first generation

and other systemic issues.

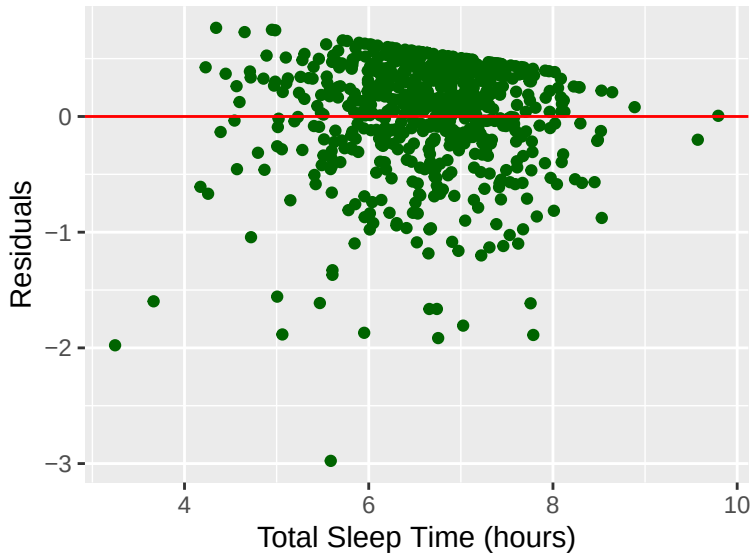

Figure 7: Total Sleep Time and Term GPA Residual Plot.

Finally, to check the error assumptions of linear regression, we check the residual plot. Figure 7 displays the residual values plotted against the predictor variable. The plot does not justify the error assumptions for linear regression. Since the points seem to follow a trend (not random scatter) and the noise is not symmetric around the line, we cannot assume the following: the mean of the errors are 0, the variance is sigma^2, and the errors are uncorrelated. This may indicate that the model has to be adjusted.

Discussion

The analysis foud that sleep time and GPA are positively associated with each other; thus there is evidence that students who sleep less indeed get lower GPAs. With 2 hours less sleep, we expect to see a decrease in GPA. Thus, it should be taken into consideration to change morning class schedules. In particular, 8AM classes are quite unpopular with students because of how early they have to get up. Since a lot of students are also staying up late, this severely reduces the amount of sleep students are getting. Seeing as the linear regression model indicates a relationship between less sleep and lower GPA, it would be worthwhile to consider either getting rid of 8AM classes or supplementing students' sleep/academics in another manner. For instance, the university can consider changing coursework limits or offering time management tips. However, it is important to note that this relationship between sleep time and GPA is not causal, meaning solely sleeping more will not cause you

to have a higher GPA. There are a myriad of reasons that could explain a casual relationship (such as demographic factors, hours spent studying, etc.), but a linear regression simply models correlation.

There are various limitations involved concerning the data set as well as the relationship models. For starters, GPA can also vary based upon a student's chosen major. Based on major, the difficulty of classes as well as grading can differ tremendously; certain majors may be getting more sleep than others, which could potentially lead to patterns in the data which reflect the trend. Additionally, the data set also includes data from three different universities, all of which could potentially have different standards of grading and thus not reflect a comparable GPA value. This can invalidate the conclusion that sleep time and GPA have a positive relationship because the data is technically inaccurate. Additionally, the residual plot revealed that there may be issues with the model used. Since there are trends present in the residual plot, we can improve this model and analysis by possibly controlling for cumulative GPA or instead modeling the change in GPA between the semesters. We can also add additional variables such as first generation status, gender, etc. (these variables are already in the data set) to further analyze trends in regards to these variables and GPA. Thus, obtaining more standardized data and controlling for additional variables can improve this analysis, however, there is statistically significant evidence from this model that sleep time has a positive relationship with GPA.