The relationship between race, gender, first-gen status, and college type for sleep and GPA in college students

Wale: Liane, Amy, Eshan, Will

2024-10-31

Introduction and Data

Project Motivation and Research Question:

As college students, we are interested in exploring how academic performance is affected differently by lack of sleep, whether a student goes to a public or private university, and more as many of these issues affect us currently. As shown in previous research, sleep impacts students' academic achievement significantly (Jalali et al. 2020), but we aim to explore this in terms of the time students went to bed, average sleep time, and more while also accounting for students' background and the type of university they go to. It is generally understood that lower levels of sleep negatively impact academic performance, but we are interested in how this impact varies or might be challenged by different factors and how we may be able to predict academic performance based on different factors. We hypothesize that the average time in bed will have the largest effect on cumulative GPA and that having less variation in bed time will lead to a higher cumulative GPA. We also anticipate the type of university students attend and first-gen status to have an affect on students' GPA. Our research question is as follows: How does sleep impact academic performance across demographics of college students?

Description of the data set and key variables:

The data was originally collected with participants being first-year students at the following three universities: Carnegie Mellon University (CMU), a STEM-focused private university, The University of Washington (UW), a large public university, and Notre Dame University (ND), a private Catholic university. To collect data on sleep, each participating student was given a device to track their sleep and physical activity for a month in the spring term, and demographic data was provided by university registrars (University 2023).

There were originally 634 observations, representing the 634 participants in this study. We filtered out students whose data was collected less than 50% of the term, leaving us with

588 participants. `demo_race`is a binary variable with 0 being underrepresented students and 1 being non-underrepresented students. Students are considered underrepresented if either parent is Black, Hispanic or Latino, Native American, or Pacific, and students are deemed non-underrepresented if both parents have White or Asian ancestry. The gender of the subject is also binary with 0 being male and 1 being female. First-generation status is binary with 0 being non-first gen and 1 being first-gen. The mean successive squared difference of bedtime measures the bedtime variability, specifically the average of the squared difference of bedtime on consecutive nights. To measure academic performance, we will be using variables `term_gpa` and `cum_gpa` (cumulative GPA) as response variables.

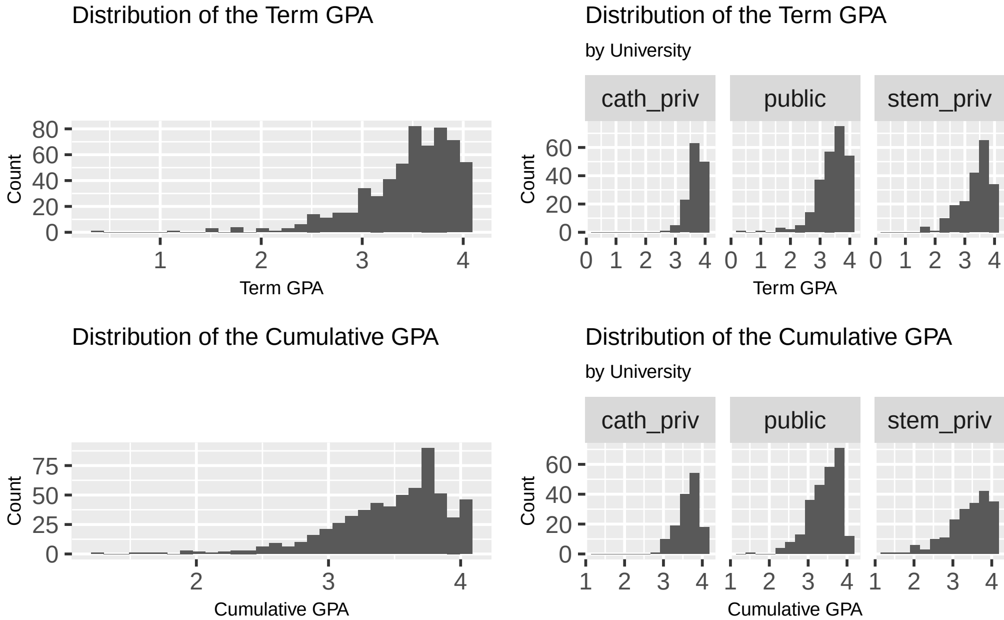
Then, we created four new variables to help with our analysis. First, we created `gpa_split` which is a binary variable that classifies GPA as "High" or "Low". A "High" GPA was determined as above the 75th percentile (3.81 GPA) of the overall term GPAs. "Low" GPA represents all the term GPAs below the 75th percentile. We then created a new variable `university`, which combines studies done at the same universities on different years ranging from 2016 to 2019. We also created the variables `threshold_gpa`, and `daytime_sleep_lvl`. `threshold_gpa` is a binary variable which classifies GPA as "high" if a student's term GPA is higher than or equal to their cumulative GPA, and "low" if it is less than their cumulative GPA. `daytime_sleep_lvl` is a binary variable that uses a threshold of 60 minutes to determine whether a student's average daytime sleep is long (high) or short (low).

Univariate EDA of The Response & Key Predictor Variables:

```
Table: Summary Statistics of Term GPA

|  Min|    Q1| Median| Mean|   Q3| Max|    SD|
|----:|-----:|------:|----:|----:|---:|-----:|
| 0.35| 3.237|  3.555| 3.45| 3.81|   4| 0.491|
```

Table 1: Summary of Cumulative GPA by University

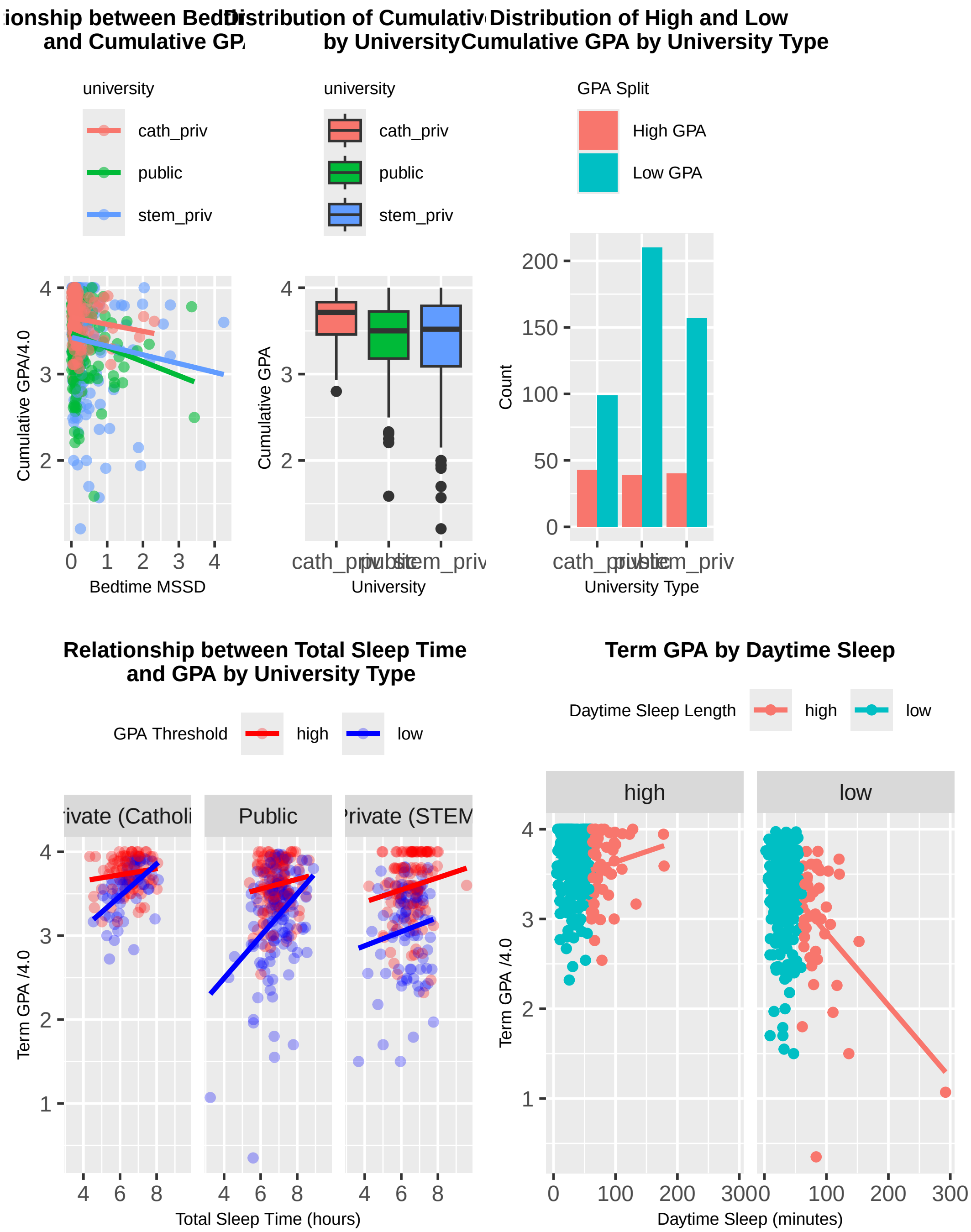
university	mean_cgpa	median_cgpa	sd_cgpa	min_cgpa	max_cgpa	count
cath_priv	3.639	3.714	0.261	2.800	4	142
public	3.429	3.501	0.400	1.588	4	249
stem_priv	3.388	3.520	0.554	1.210	4	197



The summary tables show the summary statistics for term GPA and cumulative GPA.

These four graphs show the counts of term GPA and cumulative GPA, split by university type, and all together. One notable point is that for the catholic private school and the public school, there is a very significant difference in the count of 4.0 term GPA and 4.0 cumulative GPA. This suggests that there were a number of students at these two schools who did not have a 4.0 GPA first semester, but had a 4.0 GPA second semester. This contrasts the stem private school, whose count of 4.0 GPA's for both term GPA and cumulative GPA are very close. This suggests that at the stem private school, the student's who had a 4.0 first semester are also getting a 4.0 in the second semester.

Bivariate EDA of The Response & Key Predictor Variables:



First, we looked at the relationship between total sleep time and cumulative GPA across the factors of race, gender and first-generation status. The slopes of the line of best fit for each level of each factor were all parallel, indicating that there is most likely no interaction effect between these factors with total sleep time (refer to appendix figure 1).

From the graphs above, a few of the key variables seem to have some interaction effects, and a few others do not. The first graph is a scatterplot of the relationship between total sleep time and cumulative GPA, factored by race, where red points were underrepresented students, and blue points were non-underrepresented students. The slopes of the lines best fit for each level are very similar but the slope for the underrepresented students is slightly larger than the slopes for non-underrepresented students, so there might be an interaction effect there that is worth further analysis.

The second graph, is also a scatterplot of the relationship between total sleep time and cumulative GPA, but instead factored by gender, where the red points represent male gender and the blue points represent female gender. The slopes for the line best fit for each level were essentially the same, so there is no obvious interaction effect in this graph that is worth further analysis.

The third graph shows the relationship between a student's term GPA and their total sleep time, but is facet wrapped by the university the student attended. A fourth variable, `threshold_gpa`, is a factor of 0 and 1, where 0 represents that the student's term GPA is greater than or equal to their cumulative GPA, and 1 represents that the student's term GPA is less than their cumulative GPA. This essentially tells us whether the student's term GPA is better or worse than their average GPA. Since this study only collected data during the singular term, this variable will help us determine whether a student with a low term GPA relative to their cumulative GPA is predictive of that student's total sleep time. There are a few interesting things to note of this graph. First, the term GPA of students at the STEM university seem to be more variable than the other two universities, and the total sleep time of the students at the STEM university seem to be on average lower than the other two universities.

In regards to the interaction effects, it seems as if for all three universities there is an interaction effect between students whose term GPA is less than their cumulative and student's whose term GPA is greater than or equal to their cumulative GPA. We assume this, because for all three universities, we fit a line best fit to for both term GPA $<$ cumulative GPA and vise versa, and the slopes of both lines for all three universities are different. Most notably, for the private catholic university and the public university, the slopes of the level for term GPA $<$ cumulative GPA is greater than the slopes of the level for term GPA $\geq$ cumulative GPA. This means that there is a potential interaction effect that could be explored further.

Another graph with another potential interaction effect is the sixth graph, which plots the relationship between the mean successive squared difference of bedtimes (bedtime_mssd) and a student's cumulative GPA. The points on this scatterplot were differentiated by university, with red representing the catholic private university, green representing the public university, and blue representing the STEM private university. We fit the line best fit for each of these levels, and the slope of the line for the catholic private university and the stem private university were essentially the same, but the slope of the line for the public university was slightly smaller, which means there could be a potential interaction effect there that is worth further exploration.

Methodology

Our general thought process to try and model out an answer to our research question was to think about GPA as our response variable as stated above. We felt logistic regression was a better choice using our transformed binary gpa variable, gpa_split, where a GPA above 3.0 was considered "High", and below was considered "Low." We were more concerned with understanding and predicting general ranges of academic performance as opposed to a certain GPA mark. To begin, we fit a logistic regression model with all of the variables we spoke of above (demographic variables, sleep-related variables, and a performance variable (threshold GPA)).

term	estimate	std.error	statistic	p.value
(Intercept)	3.435	1.095	3.138	0.002
TotalSleepTime	-0.003	0.002	-1.254	0.210
universitycath_priv	-0.774	0.278	-2.784	0.005
universitystem_priv	-0.284	0.272	-1.044	0.297
daytime_sleep_lvllow	-0.161	0.331	-0.488	0.626
demo_firstgen	1.130	0.423	2.672	0.008
demo_gender	0.165	0.216	0.762	0.446
bedtime_mssd	0.593	0.348	1.701	0.089
demo_race	-0.668	0.329	-2.033	0.042
threshold_gpalow	-0.505	0.214	-2.365	0.018

We saw that (by p-value), some of these predictors were not considered significant, and so we needed to do some more analysis to figure out what was necessary to use in the final model.

Analysis of Deviance Table

Model 1: as.factor(gpa_split) ~ university + demo_race + threshold_gpa

Model 2: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime

term	estimate	std.error	statistic	p.value
(Intercept)	2.623	0.357	7.339	0.000
universitycath_priv	-0.931	0.258	-3.612	0.000
universitystem_priv	-0.334	0.252	-1.329	0.184
demo_race	-0.804	0.319	-2.519	0.012
threshold_gpalow	-0.461	0.210	-2.200	0.028

```
Analysis of Deviance Table
```

```
Model 1: as.factor(gpa_split) ~ university + demo_race + threshold_gpa
Model 2: as.factor(gpa_split) ~ university + demo_race + threshold_gpa +
    TotalSleepTime
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       582     576.27
2       581     570.87  1   5.3926  0.02022 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We then created a new model, sig_fit, that isolated only the variables that were significant in the output from our original model. However, there were certain variables that we felt were still necessary to assess further, and so we used a Drop-in Deviance test to compare two models, the exact same, except one included TotalSleepTime, and the other didn't. Given TotalSleepTime being central to our entire research motivation, we wanted to investigate further, and our anova table showed us the p-value being 0.02, meaning its inclusion significantly improves the fit of the model.

Analysis of Deviance Table

Model 1: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime

Model 2: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime + bedtime_mssd

```
  Df Deviance Pr(>Chi)
1
2  1    3.453  0.06314 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0.06, greater than the threshold of 0.05, we can conclude that the inclusion of bedtime_mssd does not significantly improve the model fit and don't include it in our final model.

Analysis of Deviance Table

Model 1: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime

Model 2: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime + demo_firstgen

```
  Df Deviance Pr(>Chi)
1
2  1   8.2432  0.00409 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0.00409, we can conclude that the inclusion of demo_firstgen significantly improves the model fit and should be included in our final model.

Analysis of Deviance Table

Model 1: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime + demo_firstgen

Model 2: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime + demo_firstgen + daytime_sleep_lvl

```
  Df Deviance Pr(>Chi)
1
2  1  0.59806   0.4393
```

With a p-value of 0.4, which is greater than the threshold, we fail to reject the null hypothesis and can conclude that the inclusion of daytime_sleep_lvl does not significantly improve the model and do not include it in our final model.

We then decided to check certain interaction effects between significant main effects due to graphs in EDA (or figures in appendix?) ** FIX THIS

Interaction Effect Analyses:

Analysis of Deviance Table

Model 1: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime + demo_firstgen

Model 2: as.factor(gpa_split) ~ university + demo_race + threshold_gpa + TotalSleepTime + demo_firstgen + university*TotalSleepTime

```
  Df Deviance Pr(>Chi)
1
2  2  0.38291   0.8258
```

With a p-value of 0.8 which is greater than the threshold of 0.05, we can conclude that the interaction effects between university and TotalSleepTime are not significant enough to be included in the final model.

```
  Df Deviance Pr(>Chi)
1
2  2   1.2348   0.5393
```

With a p-value of 0.5, greater than the threshold of 0.05, we can conclude that the interaction effects between university and threshold GPA are not significant enough to be included in the final model.

We then decided to check for multicollinearity given the interconnected nature of some of the variables. We had to use GVIF because there are a few categorical predictors used.

```
                   GVIF Df GVIF^(1/(2*Df))
university     1.157534  2        1.037250
demo_race      1.014778  1        1.007362
threshold_gpa  1.017858  1        1.008889
TotalSleepTime 1.082067  1        1.040224
demo_firstgen  1.060120  1        1.029621
```

DID NOT INCLUDE INTERACTION TERMS TO SEE WHAT MAIN EFFECTS MIGHT BE MULTICOLLINEARITY

For our final model, the GVIFs (adjusted) for all variables are not greater than 10 and are very close to 1, so we can confidently assume no multicollinearity.

Results

The final model we determined is:

EDIT THESE WITH FINAL MODELS AFTER MODEL IS FINALIZED!!

$$logit(p_{high\_gpa}) = 4.206 - 2.825 \times universitycath\_priv + 0.276 \times universitystem\_priv - 1.049 \times demo\_race - 1.036$$

$$p_{high\_gpa} = \frac{1}{1 + e^{2.544 - 2.825 \times universitycath\_priv + 0.276 \times universitystem\_priv - 1.049 \times demo\_race - 1.036 \times threshold\_gpalow - 0.00}}$$

term	estimate	std.error	statistic	p.value
(Intercept)	4.206	1.020	4.124	0.000
universitycath_priv	-0.871	0.274	-3.173	0.002

term	estimate	std.error	statistic	p.value
universitystem_priv	-0.286	0.266	-1.076	0.282
demo_race	-0.694	0.325	-2.135	0.033
threshold_gpalow	-0.475	0.212	-2.238	0.025
TotalSleepTime	-0.005	0.002	-2.042	0.041
demo_firstgen	1.088	0.422	2.580	0.010

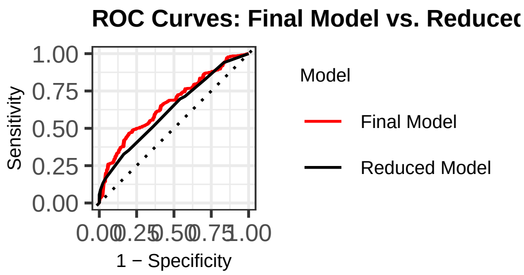
To confirm that the final model with predictors `university`, `demo_race`, `threshold_gpa`, `TotalSleepTime`, and `daytime_sleep_lvl` is better for predicting a high or low GPA (`gpa_split`) than the reduced model with initial significant predictors `university`, `demo_race`, and `threshold_gpa`, ROC and AUC were calculated for both models and compared.

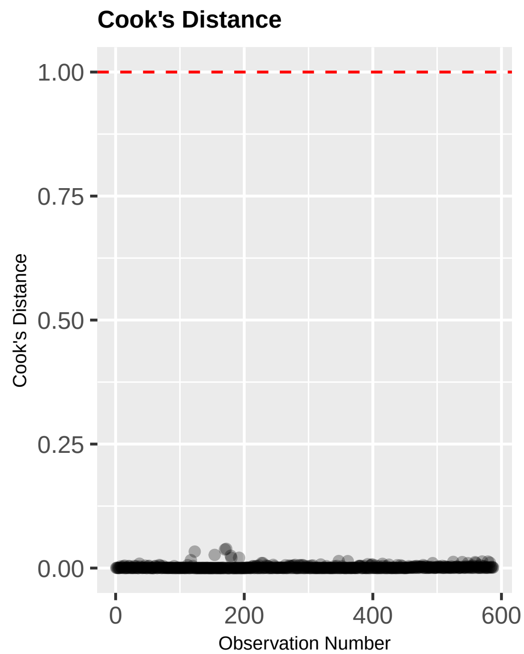
The final model we chose showed a larger AUC. The area under the curve for the final model is 0.778, whereas for the reduced model it is 0.75, showing that this final model maximizes sensitivity, the True Positive Rate, and minimizes 1 - specificity, the False Positive Rate, slightly better than the reduced model.

```
AUC for Reduced Model: 0.6250044
```

```
AUC for Final Model: 0.6644104
```

```
# A tibble: 0 x 2
# i 2 variables: obs <int>, cooks_d <dbl>
```





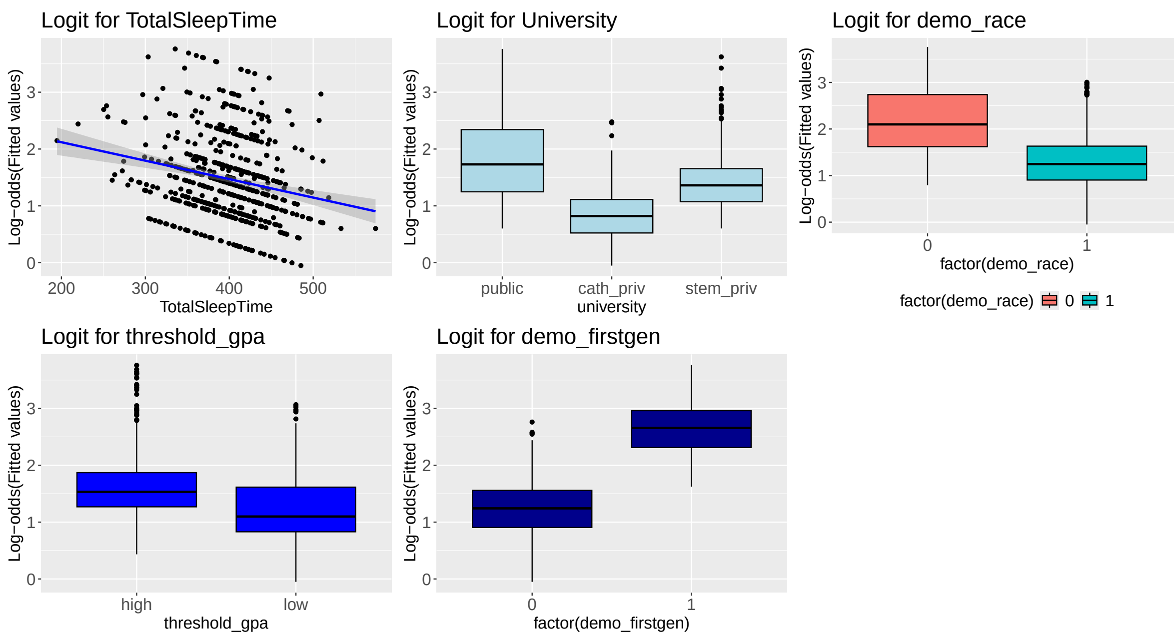
When checking for Cook's Distance, no data points were found to have a Cook's Distance greater than 1 with most far below 1, indicating that there are no influential points.

We also checked AIC and BIC for the reduced and final models:

```
AIC for Reduced Model: 586.2663

AIC for Final Model: 576.6305

BIC for Reduced Model: 608.1414

BIC for Final Model: 607.2556
```

Although the BIC for the final model is higher, because the aim of this study is to determine what combination of predictors works best to predict if a student has a high or low GPA, AIC is a more appropriate gauge for determining a better model. The AIC for the final model of 398.0 is lower than the AIC for the reduced model of 406.3. Therefore, we believe that our final model is a better model to predict a high or low GPA, and the addition of predictors `TotalSleepTime` and `daytime_sleep_lvl` are significant.

Finally, we assess the key assumptions of logistic regression within our model. All predictors show a linear relationship with the log-odds:



There is also no multicollinearity between predictors included in this model as the VIFs are all far below the threshold of 10.

```
                 GVIF Df GVIF^(1/(2*Df))
university    1.157534  2        1.037250
demo_race     1.014778  1        1.007362
threshold_gpa 1.017858  1        1.008889
TotalSleepTime 1.082067 1        1.040224
demo_firstgen 1.060120  1        1.029621
```

Although logistic regression assumes independence between observations, we grouped our observations by the type of university attended, which could introduce potential correlation between observations by school. However, we continued with logistic regression for the following reasons:

- We wanted to predict a categorical response variable, high vs. low GPA, from various predictors, and find the best model (from this dataset) to do so.
- We used `university` as one of the predictor variables to account for differences between observations and it was proven to be a significant predictor of `gpa_split` through our analysis.

APPENDIX:

Figure 1:

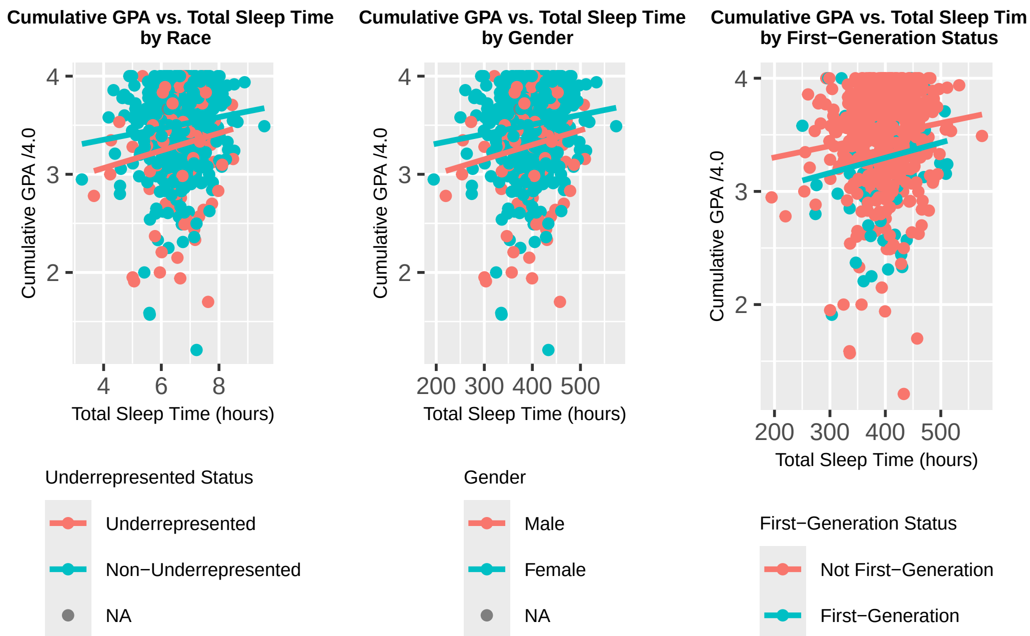


figure 2:

Table 5: Counts of NA values by University

university	total_count	na_count	non_na_count
public	249	0	249
cath_priv	142	142	0
stem_priv	197	0	197

We then looked at the confidence matrix of our model:

```
          actual
Predicted  High GPA Low GPA
  High GPA        1       1
  Low GPA       121     464
```

$$accuracy = (507 + 7)/(507 + 3 + 70 + 7) = 0.8756$$

$$misclassification = (3 + 70)/(507 + 3 + 70 + 7) = 0.1244$$

The accuracy of this model with a classification threshold of 0.5 is 0.8756. The misclassification rate of this model with a classification threshold of 0.5 is 0.1244.

Therefore, our model does well in predicting a high or low GPA with our chosen predictors.

In conclusion, significant factors that impact academic performance are university, underrepresented status, academic performance compared across semesters, total sleep time, and whether or not students napped for over 1 hour during the day.

Jalali, Rostam, Habibollah Khazaei, Behnam Khaledi Paveh, Zinab Hayrani, and Lida Menati. 2020. "The Effect of Sleep Quality on Students' Academic Achievement." *Advances in Medical Education and Practice* 11. https://doi.org/10.2147/AMEP.S261525.

University, Carnegie Mellon. 2023. "CMU Sleep Study: The Role of Sleep in Student Well-Being." https://cmustatistics.github.io/data-repository/psychology/cmu-sleep.html.