The Impact of Race, Gender, First-gen Status, Sleep Amount, and Institution Type on the Academic Performance of College Students

Wale: Liane, Amy, Eshan, Will

2024-10-31

Introduction and Data

Project Motivation and Research Question

Given our age, we are interested in exploring how collegiate academic performance is explained by various different factors, included, but not limited to, differences in a student's university type, sleep levels, and demographic background. 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: What factors affect academic performance (in terms of GPA) of college students?

Dataset 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 of years 2016 to 2019, 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. `demo_gender` and `demo_firstgen` are also both binary variables with 0 being male and 1 being female, and with 0 being non-first gen and 1 being first-gen, respectively. The mean successive squared difference of bedtime (`bedtime_mssd`) 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. The cumulative GPA is the GPA of each student's freshman fall semester.

Four new variables were created to aid our analysis. `gpa_split` was created as our response variable, 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. `university` is a variable which groups studies done by university. `threshold_gpa` is a binary variable that 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 Exploratory Data Analysis

Table 1: Summary Statistics of Term GPA

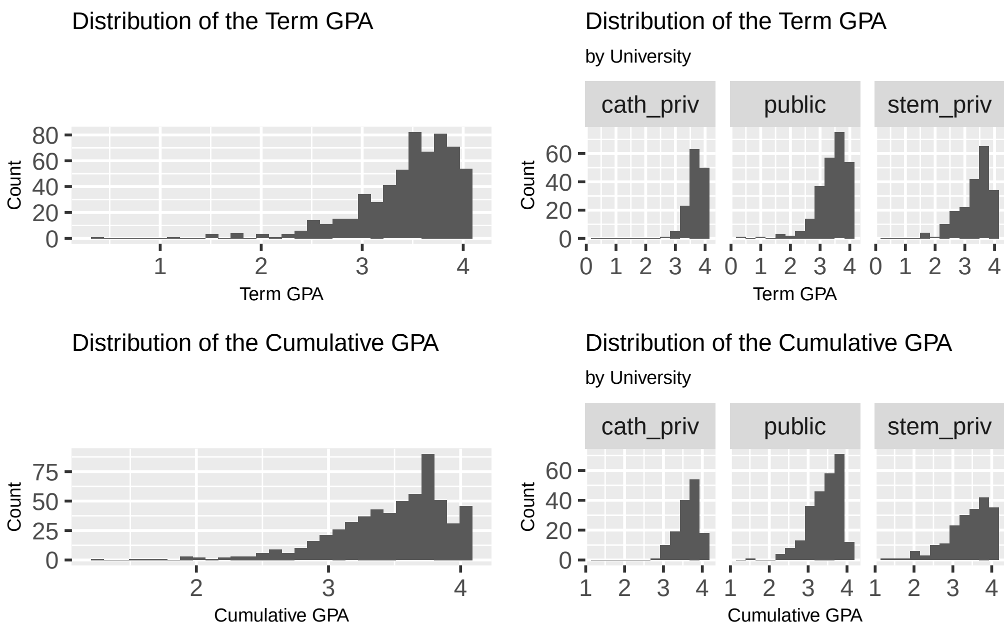
Min	Q1	Median	Mean	Q3	Max	SD
0.35	3.237	3.555	3.45	3.81	4	0.491

Table 2: 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

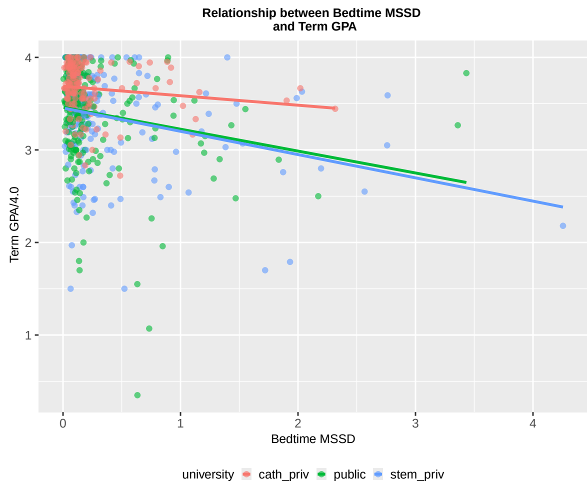
The summary table on the left shows the summary statistics for term GPA. The top 25% of students had term GPAs above 3.81, and the bottom 25% had term GPAs below 3.24. This suggests that the majority of students are doing well in school.

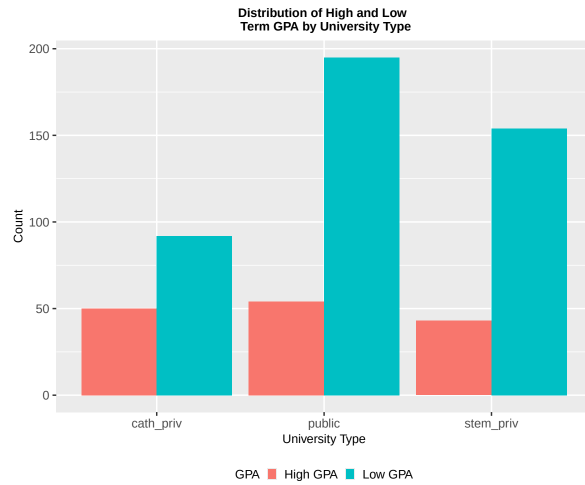
The summary table on the right shows the summary statistics for cumulative GPA. One interesting piece of information is the minimum GPA from the catholic private school is significantly higher than the stem private school and the public school. The catholic private school's mean and median cumulative GPA are also higher than the other two schools. This could suggest that the catholic private school has higher grade inflation than the other two schools.



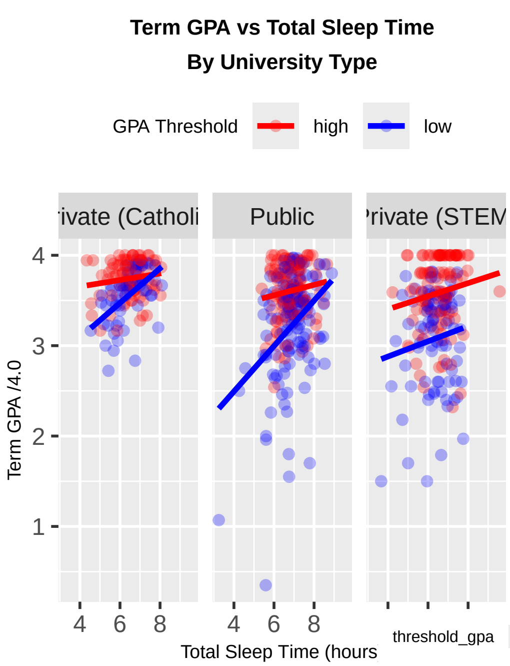
These four graphs show the counts of term GPA and cumulative GPA, split by university type, and all together. One notable point is that for the catholic private school and the public school, there is a very significant difference in the count of 4.0 term GPA and 4.0 cumulative GPA. This suggests that 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.

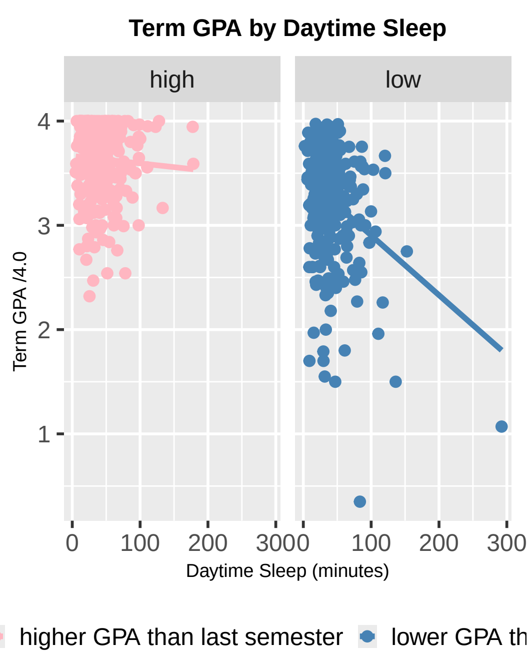
Bivariate Exploratory Data Analysis





The graph on the left shows the relationship between the mean successive squared difference of bedtimes (bedtime_mssd) and a student's cumulative GPA. The graph is colored by university type, with colors indicated on the legend above. The lines of best fit for catholic and stem private had similar slopes, however the public school slope was different, indicating the possibility of an interaction effect. The graph on the right shows each university's distribution of "High" and "Low" term GPA. The ratio between "High" and "Low" term GPA at the catholic private school is the smallest of the three universities. This is somewhat surprising, given that the mean and median cumulative GPA for the catholic private school was the highest of the three.





The graph on the left shows the relationship between a student's term GPA and their total sleep time, but is separated by university. 'threshold_gpa' 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 across universities. It seems that there is a potential interaction effect between 'gpa_threshold' and 'university' since the relationship between Term GPA and Total Sleep Time is impacted by both 'gpa_threshold' and university with varying patterns of 'gpa_threshold' for all schools. The graph on the right shows the relationship between 'term_gpa' and 'daytime_sleep' by 'gpa_threshold'. We notice that if you had a higher term GPA than cumulative GPA, taking more naps did not significantly impact GPA on average. But for those that had a lower term GPA than cumulative GPA, taking more naps significantly reduced GPA on average. This suggests a potential interaction effect between 'gpa_threshold' and 'daytime_sleep'.

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)	3.708	1.110	3.340	0.001
TotalSleepTime	-0.007	0.002	-2.774	0.006
universitycath_priv	-0.555	0.273	-2.034	0.042
universitystem_priv	0.196	0.267	0.734	0.463
daytime_sleep_lvllow	-0.123	0.311	-0.395	0.693
demo_firstgen	1.024	0.365	2.804	0.005
demo_gender	0.201	0.216	0.932	0.352
bedtime_mssd	0.412	0.340	1.214	0.225
demo_race	-0.805	0.317	-2.539	0.011
threshold_gpalow	1.740	0.239	7.268	0.000

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 + TotalSleepTime + demo_firstgen`

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

term	estimate	std.error	statistic	p.value
(Intercept)	4.187	1.054	3.971	0.000
universitycath_priv	-0.619	0.270	-2.295	0.022
universitystem_priv	0.188	0.263	0.717	0.473
demo_race	-0.828	0.314	-2.634	0.008
threshold_gpalow	1.750	0.238	7.354	0.000
TotalSleepTime	-0.008	0.002	-3.231	0.001
demo_firstgen	1.006	0.365	2.755	0.006

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
580	556.374	NA	NA	NA
579	554.584	1	1.79	0.181

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 bedtime_mssd, and the other didn't. With a p-value of 0.181, greater than the threshold of 0.05, we can conclude that the inclusion of `bedtime_mssd` does not significantly improve the model fit, so we will not include it 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_sleeplvl`

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
580	556.374	NA	NA	NA
579	556.092	1	0.282	0.595

With a p-value of 0.595, greater than the threshold of 0.05, we can conclude that the inclusion of `bedtime_mssd`daytime_sleep_lvl‘ does not significantly improve the model fit, so we will not include it in our final model.

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.

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`

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
580	556.374	NA	NA	NA
578	555.403	2	0.971	0.615

With a p-value of 0.615 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.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
580	556.374	NA	NA	NA
578	550.965	2	5.409	0.067

With a p-value of 0.067, 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	1.163	2	1.038
demo_race	1.025	1	1.012
threshold_gpa	1.027	1	1.013
TotalSleepTime	1.089	1	1.044
demo_firstgen	1.075	1	1.037

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!!

$$logit(p_{high\_gpa}) = 4.187 - 0.619 \times universitycath\_priv + 0.188 \times universitystem\_priv - 0.828 \times demo\_race + 1.750$$

$$p_{high\_gpa} = \frac{1}{1 + e^x}$$

Where x represents the logit equation shown above.

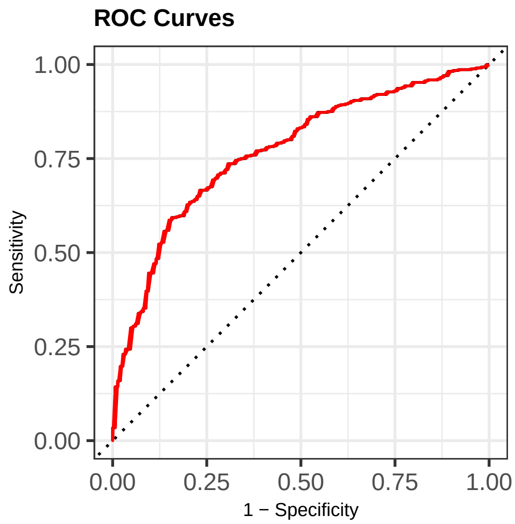
term	estimate	std.error	statistic	p.value
(Intercept)	4.187	1.054	3.971	0.000
universitycath_priv	-0.619	0.270	-2.295	0.022
universitystem_priv	0.188	0.263	0.717	0.473
demo_race	-0.828	0.314	-2.634	0.008
threshold_gpalow	1.750	0.238	7.354	0.000
TotalSleepTime	-0.008	0.002	-3.231	0.001
demo_firstgen	1.006	0.365	2.755	0.006

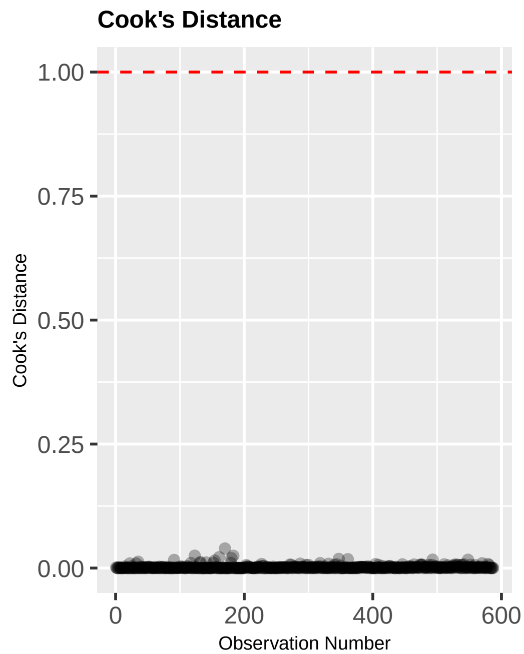
To confirm that the final model with predictors `university`, `demo_race`, `threshold_gpa`, `TotalSleepTime`, and `daytime_sleep_lvl` is better for predicting a high or low GPA (`gpa_split`) than the reduced model with initial significant predictors `university`, `demo_race`, and `threshold_gpa`, ROC 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 Final Model: 0.7688312
```

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

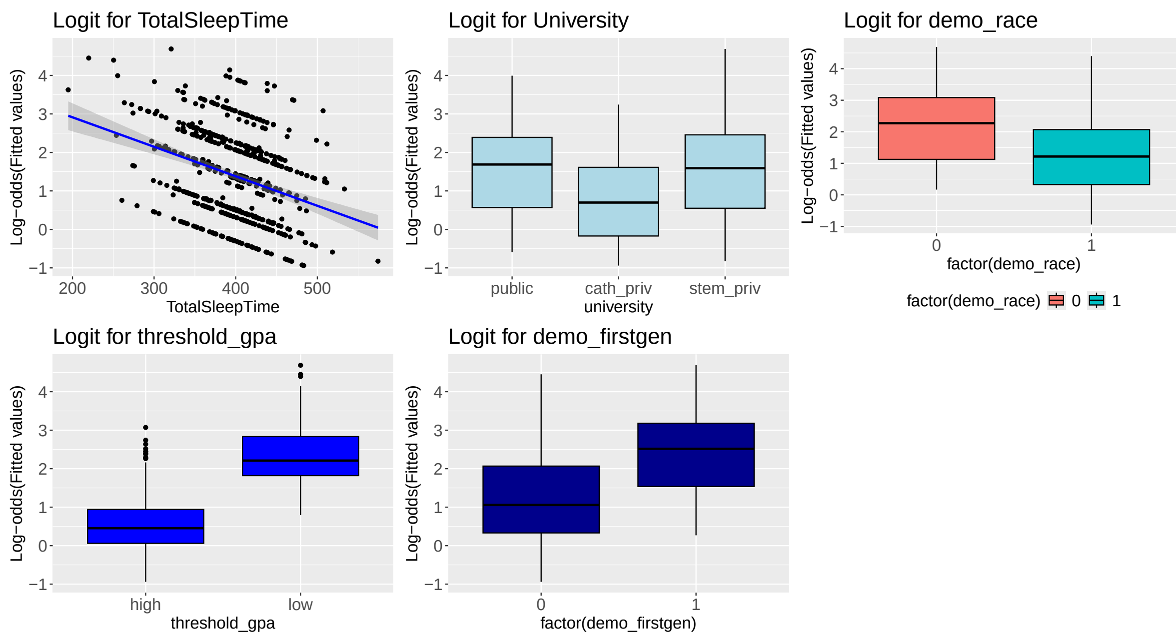




When checking for Cook's Distance, no data points were found to have a Cook's Distance greater than 1 with most far below 1, indicating that there are no influential points.

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

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	1.163	2	1.038
demo_race	1.025	1	1.012
threshold_gpa	1.027	1	1.013
TotalSleepTime	1.089	1	1.044
demo_firstgen	1.075	1	1.037

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.

Discussion and Conclusion

Factors that have a significant impact on students' academic performance include university type, race, first-generation status, total sleep time, and whether students did better in the spring semester compared to the fall semester. In contrast, factors such as bedtime variability, gender, and the amount of daytime sleep were not significant in determining students' academic performance. This gives us a full picture of how to characterize an individual (demographic and sleep-related variables). We came to this result from looking at p-values in a logistic regression model, and from using drop-in deviance tests to assess the value of adding certain variables to the model, including interaction effects.

Some limitations of this analysis are as follows:

Our cleaned dataset has a total of 588 observations, with varying numbers of students across universities, with 142 observations from the Catholic private university, 249 from the public university, and 197 from the STEM private university. Additionally, each university type (catholic private, public, and STEM private) only consisted of one university for each type (Notre Dame, University of Washington, and Carnegie Mellon respectively), which makes it difficult to generalize our conclusions to all college students across the United States. However, it is still interesting to see the comparison across university types.

Another limitation is that using term GPA as our response variable does not adequately capture all aspects of academic performance. Term GPA is only one indicator of academic performance. There are other indicators such as joining student life organizations, participating in research, internships, study abroad trips etc. which are hard to quantify. If there was more qualitative data, such as a survey that asks about a student's participation in out of the classroom activities, and a student's perceived academic performance, we could have a more holistic model that takes into account these other factors. Furthermore, GPA across universities may be calculated with different levels of grade inflation, indicating a lack of uniformity in the response variable.

Appendix

Figure 1.

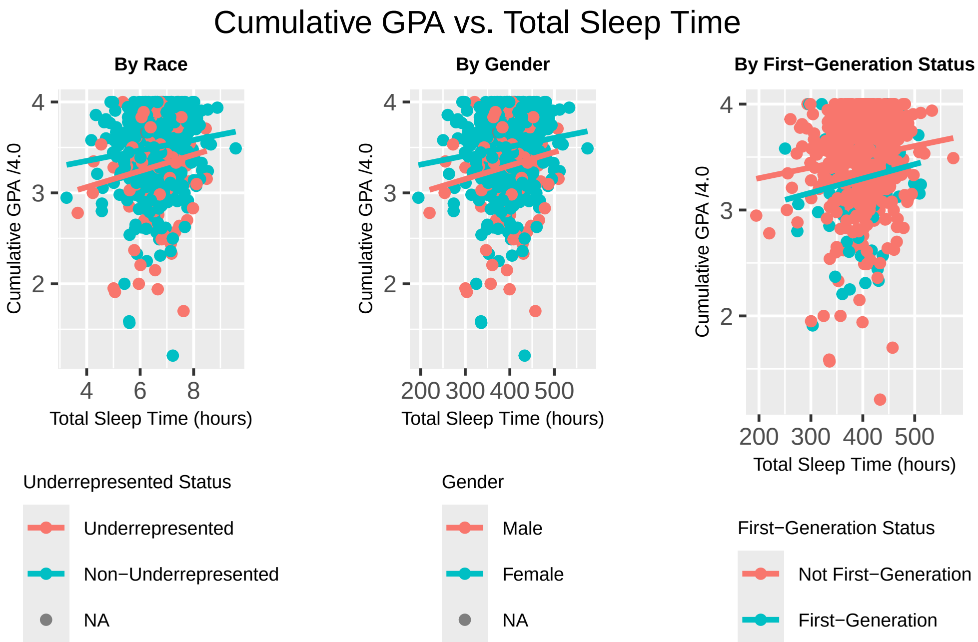


Figure 2.

Table 12: 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

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.