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

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?

Exploratory Data Analysis

Description of the data set and key variables:

The data was originally collected in 2019, with the 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 Fitbit device to track their sleep and physical activity for a month in the spring term, and grade and demographic data was provided by university registrars (University 2023).

There are 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. Race is a binary variable separated into underrepresented students and non-underrepresented students with 0 being underrepresented and 1 being non-underrepresented. 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. Furthermore, we created the variable `gpa_split` which uses a threshold of a 3.0 GPA to determine whether a student has a "low" or "high" GPA.

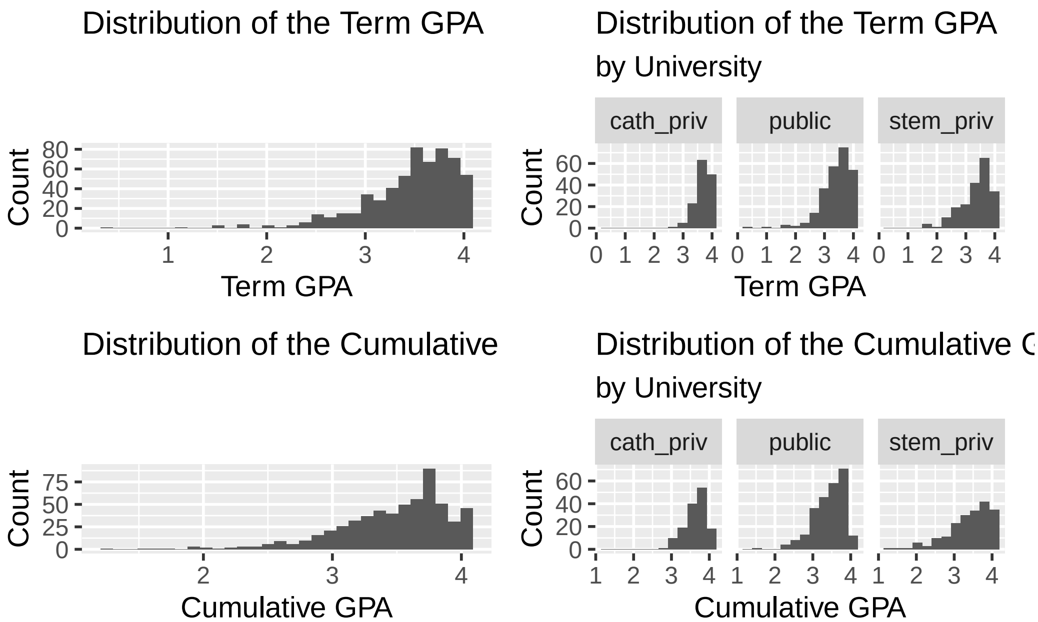
Here, we created a new variable `university`, which combines studies done at the same universities on different years ranging from 2016 to 2019. We also created the variables `gpa split` , `threshold_gpa`, and `daytime_sleep_lvl`. `gpa_split` is a binary variable which classifies a GPA as high if it is above or equal to 3.0, and low if it is below. `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:

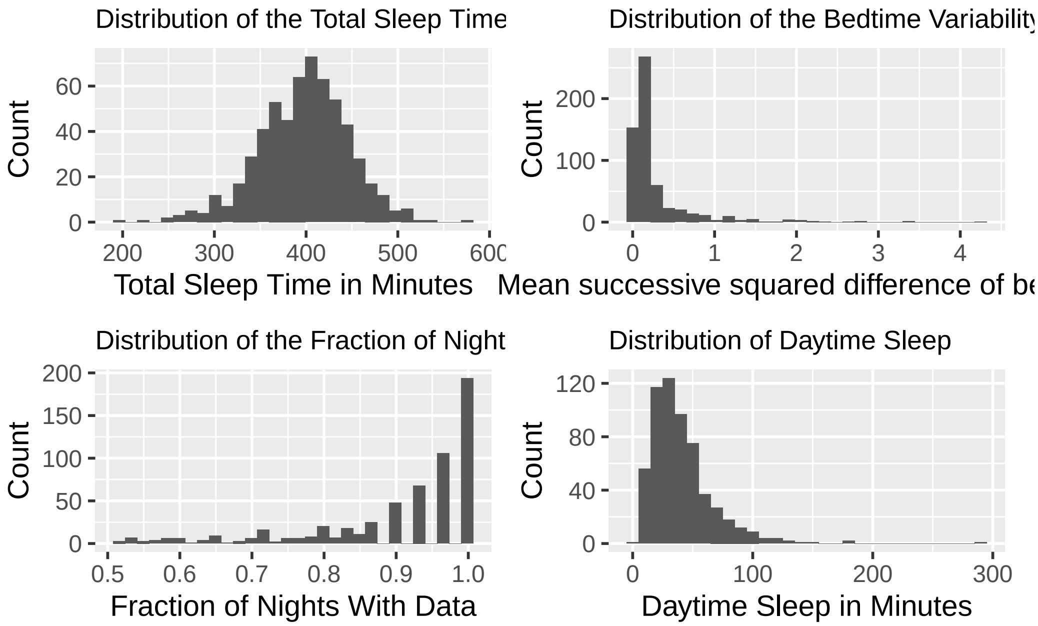
university	mean_tgpa	median_tgpa	sd_tgpa	min_tgpa	max_tgpa	count
cath_priv	3.665	3.714	0.267	2.722	4	142
public	3.401	3.500	0.518	0.350	4	249
stem_priv	3.359	3.490	0.535	1.500	4	197

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

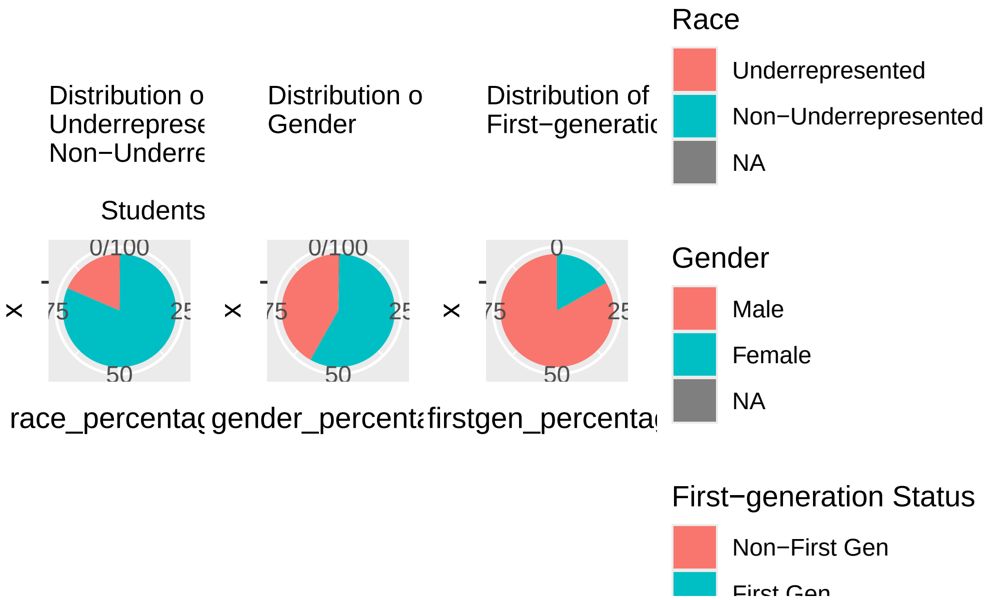
university	total_count	na_count	non_na_count
cath_priv	142	142	0
public	249	0	249
stem_priv	197	0	197



These four graphs show the counts of term GPA and cumulative GPA, split by university type, and all together. One notable points is that there are less 4.0 term GPA at the stem private school than the other two schools, but more 4.0 cumulative GPAs there than the other two schools. This could be due to the fact that classes are harder at stem private schools, so there are less 4.0 term to term than other schools, but the students may be smarter, so overall, there are more students with 4.0 cumulative GPAs. Another notable point is that the distribution of both the term GPA and cumulative GPA are unimodal and left skewed with long tails, but the term GPA has a larger spread than the cumulative GPA.

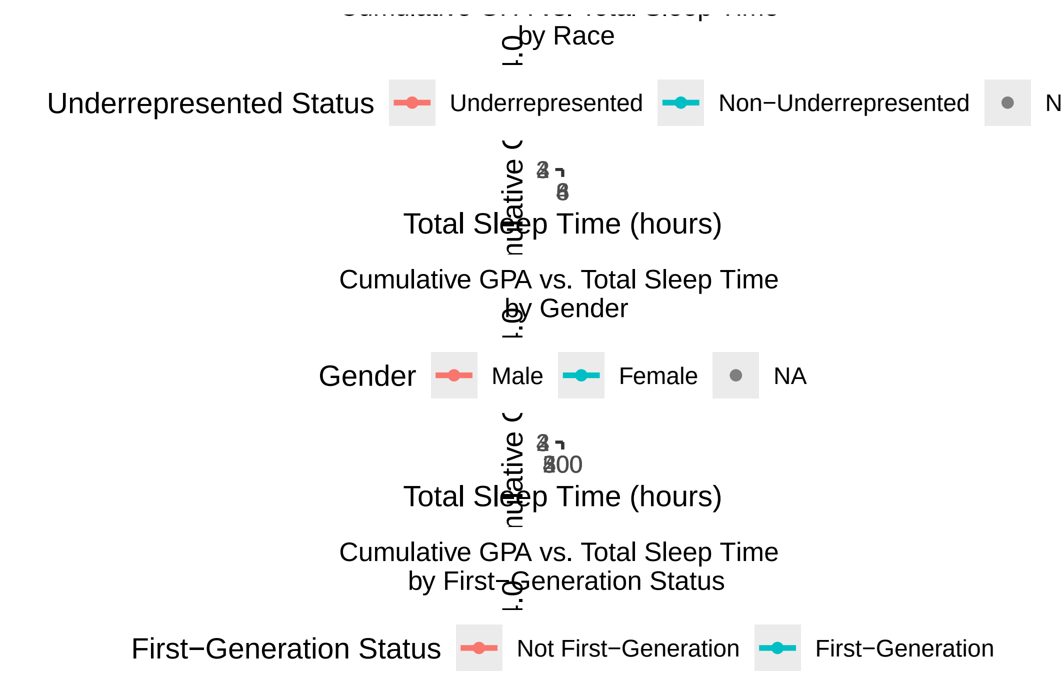


This group of graphs shows the sleep variables we used in our analysis. The top left graph displays the total sleep time in minutes, and is unimodal and symmetric. The top right graph shows the distributino of the average mean successive squared difference (MSSD) of bedtime, which is right-skewed with a long tail. This makes sense, because most people have low bedtime variability, but some people, for example people who enjoy partying in college, may have a large difference in bedtime. The bottom left graph shows the distriubtion of the fraction of nights where data was collected, after the data had been cleaned to remove values less than 50%. This shows that the majority of people collected 90% or more of their sleep data. Finally, the bottom right graphs shows the distribution of the average daytime sleep in minutes, which is unimodal right skewed. This also makes sense, because most people don't nap much, but some people enjoy routine naps.

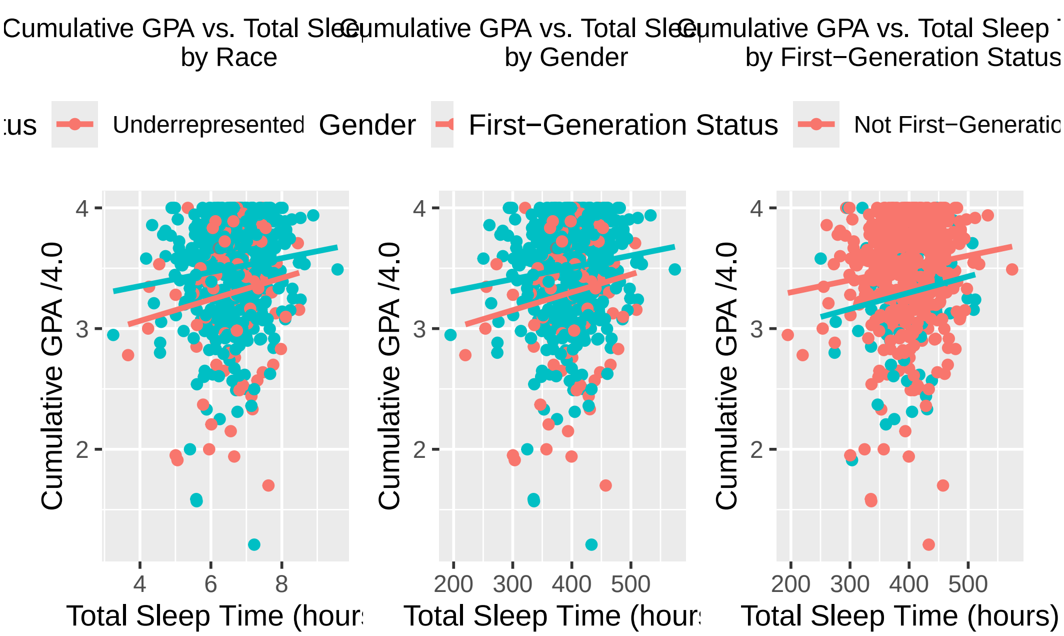


These three graphs show the distribution of difference demographics in the study. About 80% of the students are not underrepresented, around 55% identify as female, and about 20% have first generation college student status.

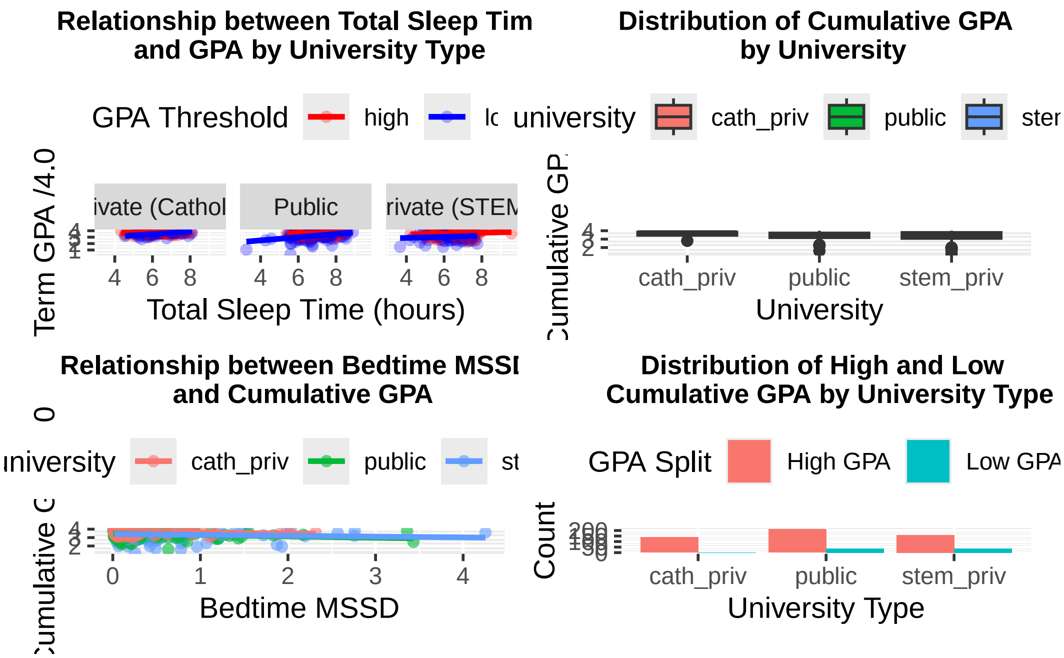
Bivariate EDA of The Response & Key Predictor Variables:

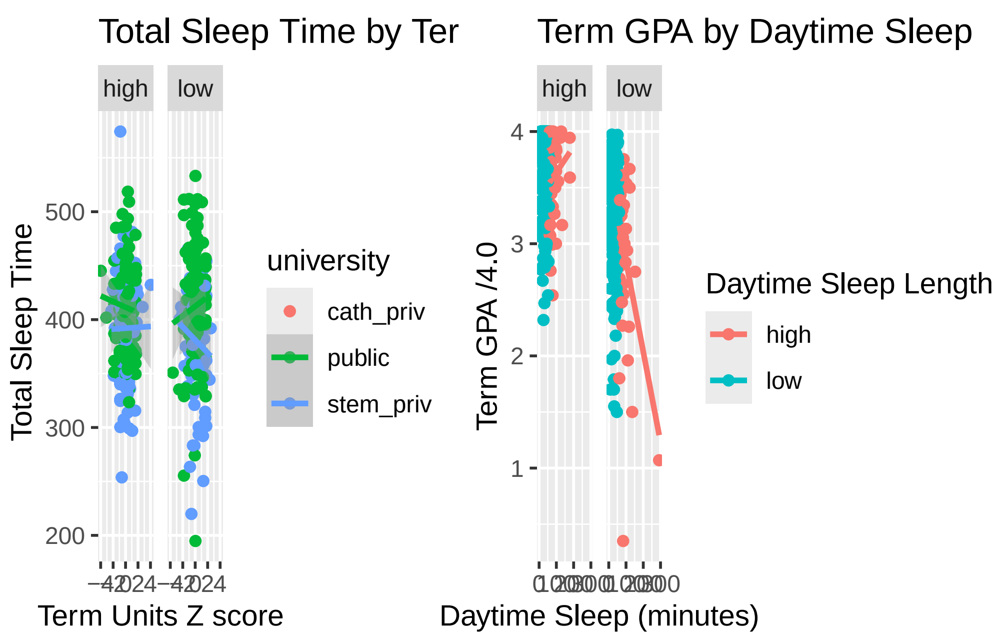


by Race
Underrepresented Status
Underrepresented
Non−Underrepresented
Total Sleep Time (hours)
Cumulative GPA vs. Total Sleep Time
by Gender
Gender
Male
Female
NA
Total Sleep Time (hours)
Cumulative GPA vs. Total Sleep Time
by First−Generation Status
First−Generation Status
Not First−Generation
First−Generation



Cumulative GPA vs. Total Slee
by Race
Underrepresented
Cumulative GPA vs. Total Slee
by Gender
Gender
Cumulative GPA vs. Total Sleep
by First−Generation Status
First−Generation Status
Cumulative GPA /4.0
4
3
2
4
6
8
200
300
400
500
Total Sleep Time (hours)





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

[INTRODUCE WHAT IS GOING ON HERE, be more details, what is "this", explain how this relates to research question, specify which predictors are not significant and its bcs pval > 0.05]

We decided that the best way to approach this is by transforming GPA into a binary variable where "High" is considered a GPA of over 3, and "Low" is below. Using this transformed response variable, we figured the best way to approach this would be to use a logistic regression model, given the binary nature of the response. We fit this model with essentially every single one of our predictors (seen above). We saw that (by p-value), some of these predictors were not significant, and so we needed to do some more analysis to figure out what was necessary to use in the model.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.702	1.345	-0.521	0.602
TotalSleepTime	-0.005	0.003	-1.859	0.063
universitypublic	2.761	0.757	3.648	0.000
universitystem_priv	3.094	0.749	4.130	0.000
daytime_sleep_lvllow	-0.589	0.325	-1.810	0.070
demo_firstgen	0.454	0.330	1.376	0.169
demo_gender	-0.490	0.269	-1.823	0.068
bedtime_mssd	0.307	0.241	1.274	0.203
demo_race	-0.878	0.306	-2.867	0.004
threshold_gpalow	-1.096	0.291	-3.769	0.000

```
# A tibble: 8 x 1
  x[,"GVIF"] [,"Df"] [,"GVIF^(1/(2*Df))"]
       <dbl>   <dbl>                <dbl>
1       1.24       1                 1.11
2       1.26       2                 1.06
3       1.12       1                 1.06
4       1.16       1                 1.08
5       1.03       1                 1.01
6       1.19       1                 1.09
7       1.08       1                 1.04
8       1.07       1                 1.04
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3.159	0.738	-4.281	0
universitypublic	2.651	0.740	3.583	0

term	estimate	std.error	statistic	p.value
universitystem_priv	2.988	0.741	4.031	0
demo_race	-1.112	0.288	-3.865	0
threshold_gpalow	-0.968	0.274	-3.528	0

term	estimate	std.error	statistic	p.value
(Intercept)	-0.227	1.268	-0.179	0.858
universitypublic	2.855	0.745	3.832	0.000
universitystem_priv	3.012	0.742	4.058	0.000
demo_race	-1.102	0.292	-3.775	0.000
threshold_gpalow	-1.042	0.280	-3.723	0.000
TotalSleepTime	-0.008	0.003	-2.806	0.005

```
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     396.29
2       581     388.33  1   7.9581 0.004787 **
---
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 0.004, meaning its inclusion significantly improves the fit of the model.

$$H_0 : \beta_{\text{TotalSleepTime}} = 0 H_a : \beta_{\text{TotalSleepTime}} \neq 0$$

term	estimate	std.error	statistic	p.value
(Intercept)	0.215	0.947	0.228	0.820
TotalSleepTime	-0.005	0.002	-2.218	0.027

term	estimate	std.error	statistic	p.value
(Intercept)	-0.227	1.268	-0.179	0.858
universitypublic	2.855	0.745	3.832	0.000
universitystem_priv	3.012	0.742	4.058	0.000
demo_race	-1.102	0.292	-3.775	0.000
threshold_gpalow	-1.042	0.280	-3.723	0.000
TotalSleepTime	-0.008	0.003	-2.806	0.005

```
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 +
    demo_firstgen + TotalSleepTime
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       581     388.33
2       580     386.05  1   2.2814   0.1309

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 +
    bedtime_mssd + TotalSleepTime
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       581     388.33
2       580     385.77  1    2.558   0.1097

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 +
    daytime_sleep_lvl + TotalSleepTime
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       581     388.33
2       580     384.03  1   4.2975  0.03817 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We then repeated this same drop-in deviance process with the "first-gen" demographic variable, the bedtime variable, and the daytime_sleep variable. The p-values generated by the first two ANOVA tables indicate that they do not significantly improve model-fit, however the inclusion of daytime_sleep_lvl does.

```
# A tibble: 4 x 1
  x[,"GVIF"] [,"Df"] [,"GVIF^(1/(2*Df))"]
       <dbl>   <dbl>                <dbl>
1       1.08       2                 1.02
2       1.02       1                 1.01
3       1.03       1                 1.01
4       1.08       1                 1.04
```

```
# A tibble: 5 x 1
  x[,"GVIF"] [,"Df"] [,"GVIF^(1/(2*Df))"]
       <dbl>   <dbl>                <dbl>
1       1.13       2                 1.03
2       1.02       1                 1.01
3       1.03       1                 1.01
4       1.10       1                 1.05
5       1.13       1                 1.06
```

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        1.134173  2        1.031977
demo_race         1.023548  1        1.011705
threshold_gpa     1.030173  1        1.014975
TotalSleepTime    1.133772  1        1.064787
daytime_sleep_lvl 1.103836  1        1.050636
```

We renamed the "model_4" from above to be the final model. The VIFs for all variables are not greater than 10 and are very close to 1, so we can confidently assume no multicollinearity.

Results

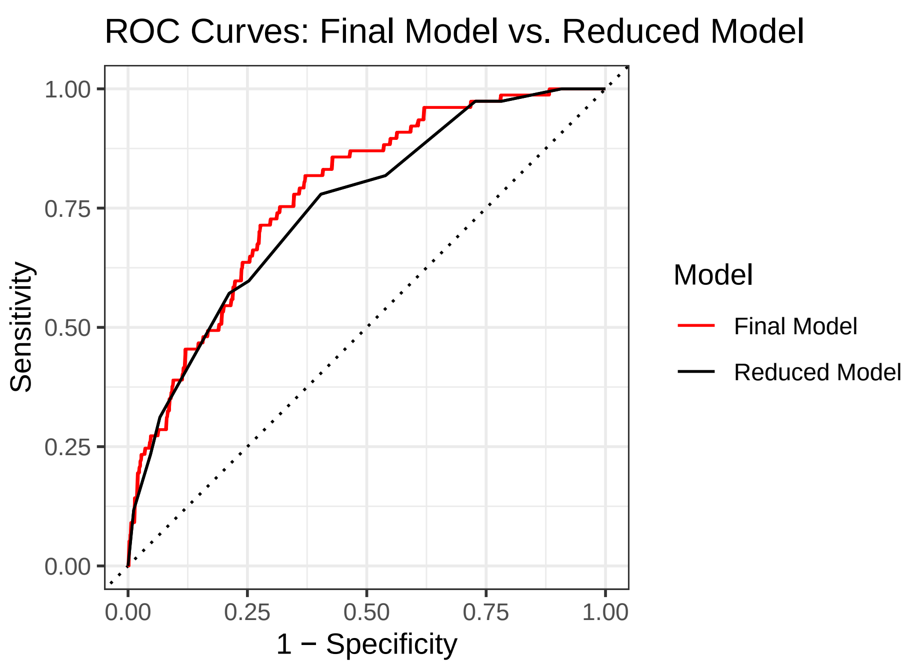
The final model we determined is:

****ADD EQUATION**

```
# A tibble: 7 x 5
  term                  estimate std.error statistic   p.value
  <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)           -0.281     1.28       -0.220 0.826
2 universitypublic       2.83      0.745       3.79  0.000150
3 universitystem_priv    3.10      0.746       4.16  0.0000322
4 demo_race             -1.05      0.296      -3.55  0.000388
5 threshold_gpalow      -1.04      0.282      -3.68  0.000236
6 TotalSleepTime        -0.00633   0.00281    -2.25  0.0243
7 daytime_sleep_lvllow -0.682     0.322      -2.12  0.0344
```

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: 0.7505093
```

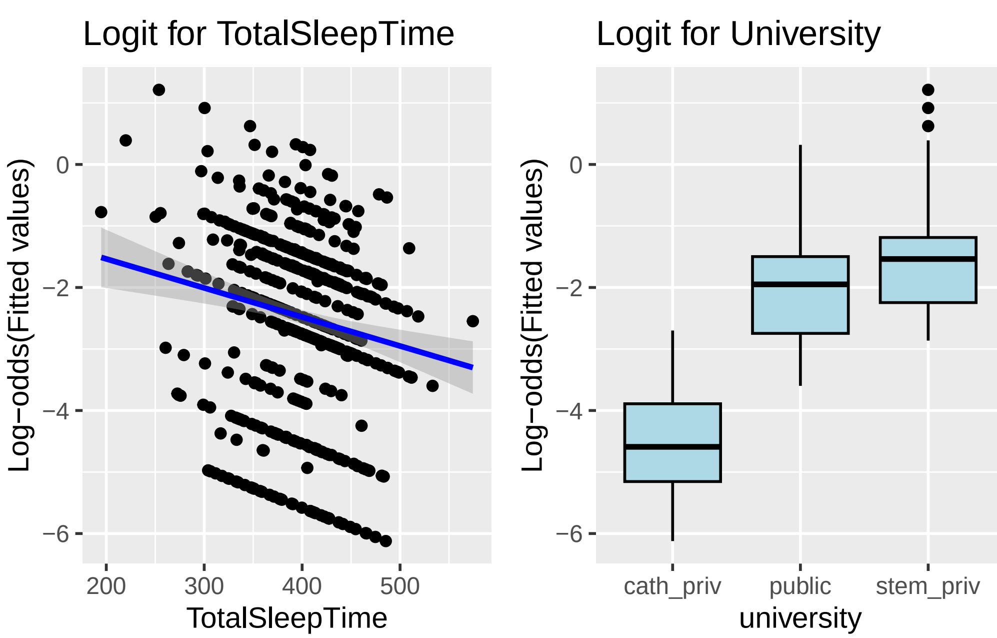
```
AUC for Final Model: 0.7782531
```

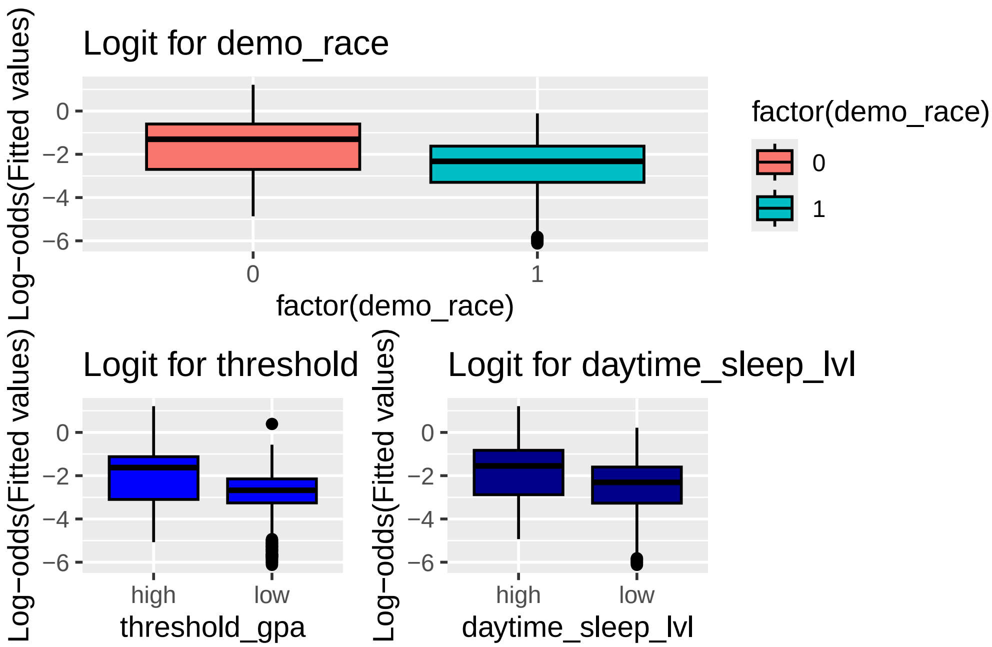
We also checked AIC and BIC for the reduced and final models:

```
AIC for Reduced Model: 406.2854

AIC for Final Model: 398.0298

BIC for Reduced Model: 428.1606

BIC for Final Model: 428.6549
```

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:





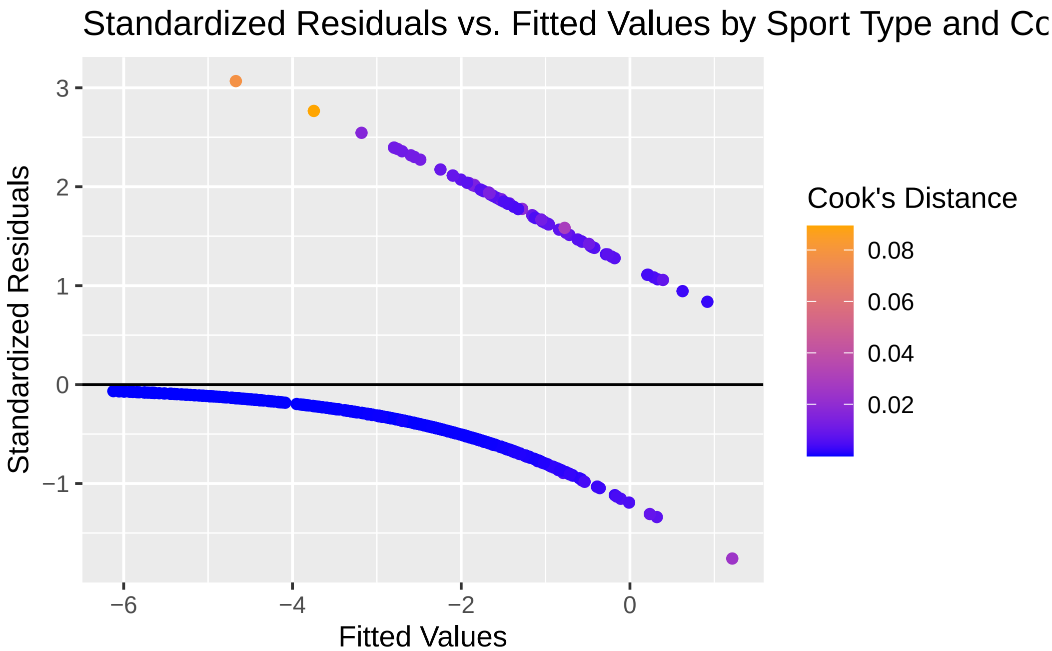
There is also no multicollinearity between predictors included in this model as the VIFs are all far below the threshold of 10.

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

```
                      GVIF Df GVIF^(1/(2*Df))
university        1.134173  2        1.031977
demo_race         1.023548  1        1.011705
threshold_gpa     1.030173  1        1.014975
TotalSleepTime    1.133772  1        1.064787
daytime_sleep_lvl 1.103836  1        1.050636
```

```
                  GVIF Df GVIF^(1/(2*Df))
university    1.018954  2        1.004705
demo_race     1.019581  1        1.009743
threshold_gpa 1.007741  1        1.003863
```

```
# A tibble: 0 x 1
# i 1 variable: cooks_d <dbl>
```



INCLUDE GRAPH OR NO? do we need fitted vs resid or ?

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.

We then looked at the confidence matrix of our model:

```
          actual
Predicted  High GPA Low GPA
  High GPA      507      70
  Low GPA         3       7
```

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

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

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.

Things that still need to be done:

Introduction and data:

Explain the univariate and bivariate EDA.

(why did we use term gpa, cum gpa and gpa_split as different response variables? how did we decide that gpa_split was the best response variable?)

Explain if there are any interaction terms/ if we should transform any variables.

Choose graphs!!

Methodology:

Explain model selection (Explain how we got to the sig_fit model from the log_model/ why did we choose log_reg

Explain how we got to the university_final model from the sig_fit model)

Included in that is:

- anova tables
- The hypothesis test / drop in deviance test results

Results:

Interpret AIC, ROC, AUC for final model, explain results/ compare to the sig_fit model.

- can't do R-sq for log reg

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.