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

Your written report goes here!

> **❗ Important**
>
> Before you submit, make sure your code chunks are turned off with `echo: false` and there are no warnings or messages with `warning: false` and `message: false` in the YAML.

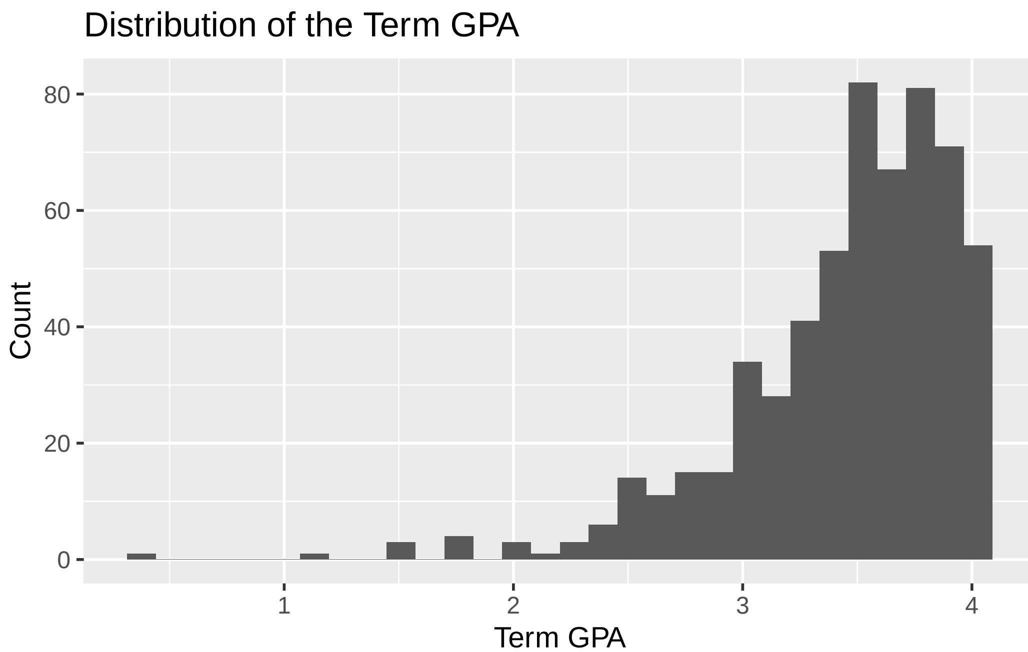
Exploratory Data Analysis

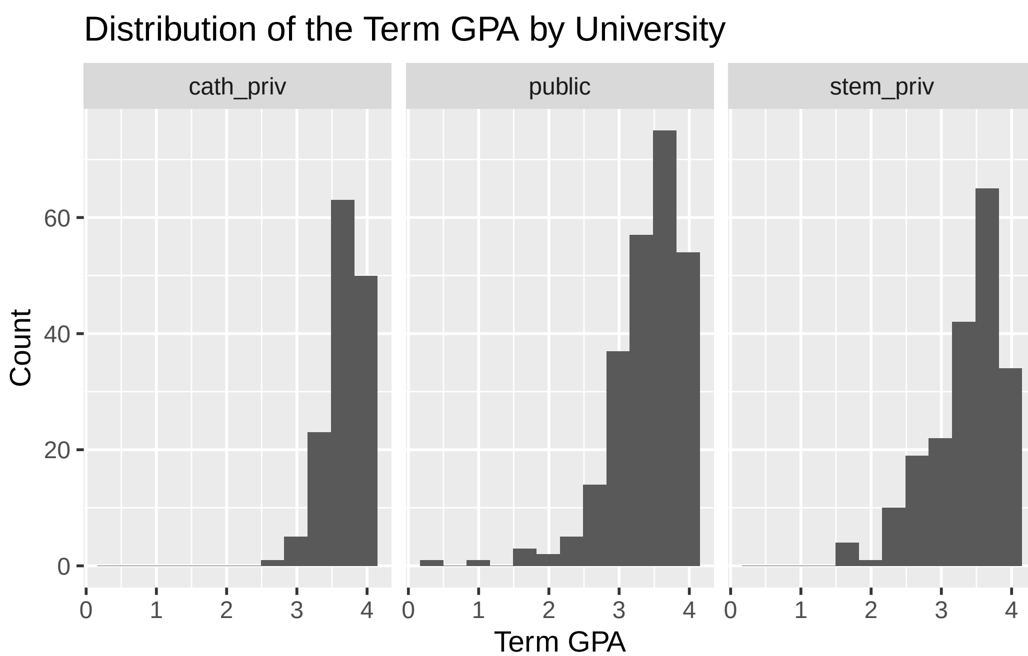
Description of the data set and key variables.

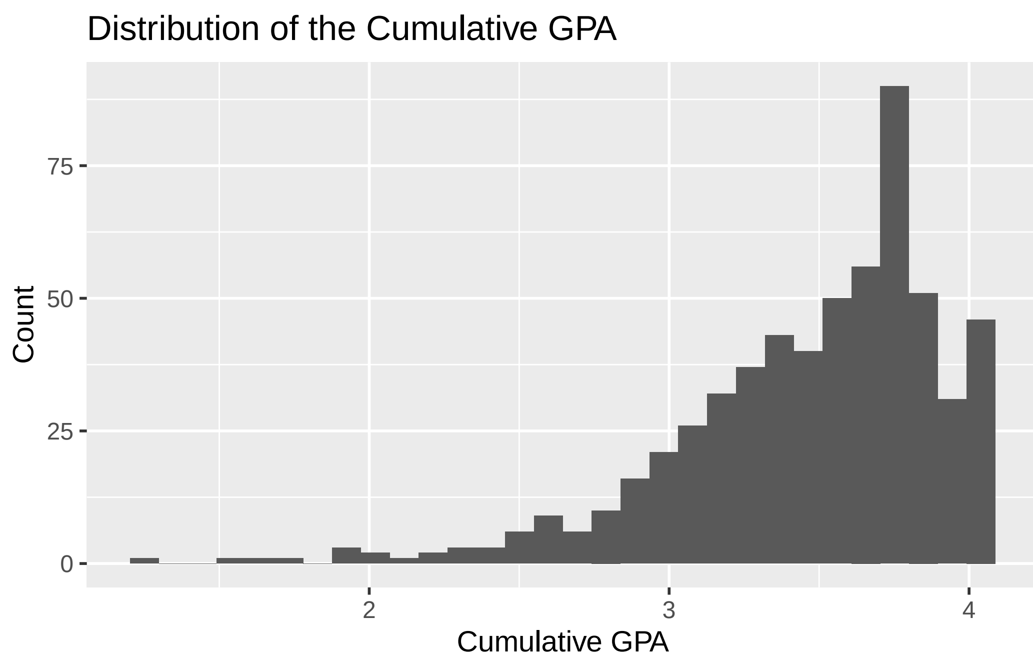
The data was originally collected in 2019, with the participants being first-year students at the following three universities: Carnegie Mellon University (CMU), a STEM-focused private university, The University of Washington (UW), a large public university, and Notre Dame University (ND), a private 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.

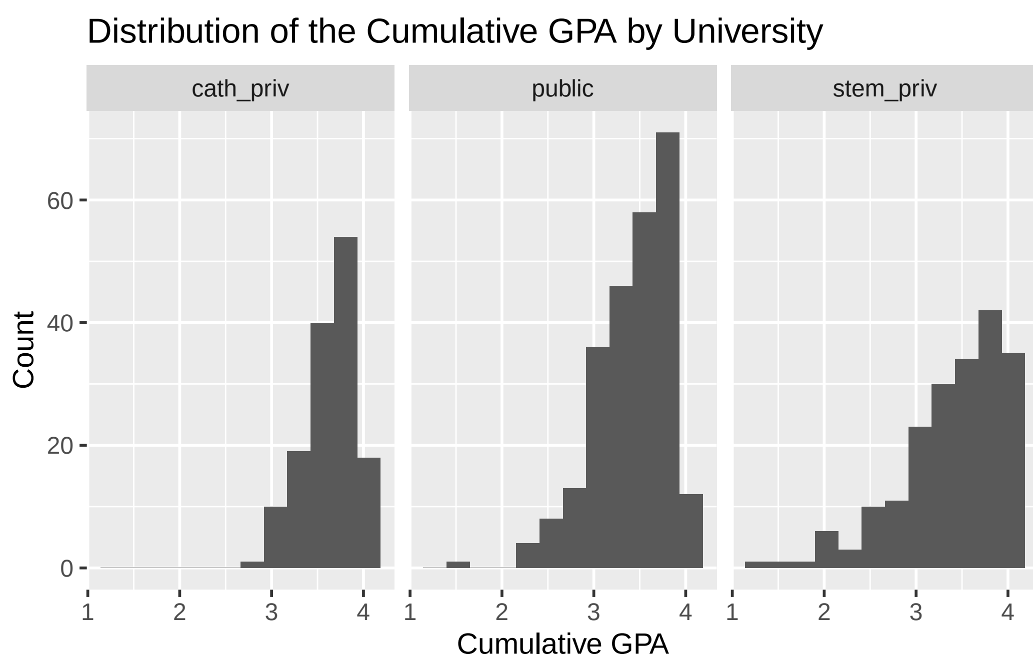
There are 634 observations, representing the 634 participants in this study. Race is a binary variable separated into underrepresented students and non-underrepresented students with 0 being underrepresented and 1 being non-underrepresented. Students are considered underrepresented if either parent is 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.

Univariate EDA of The Response & Key Predictor Variables:





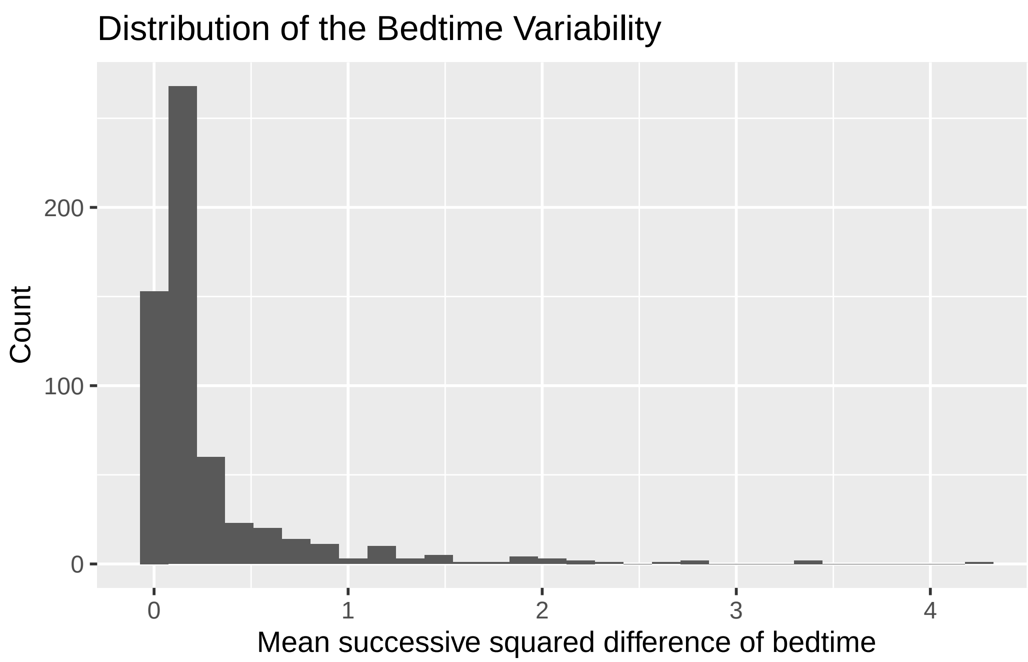




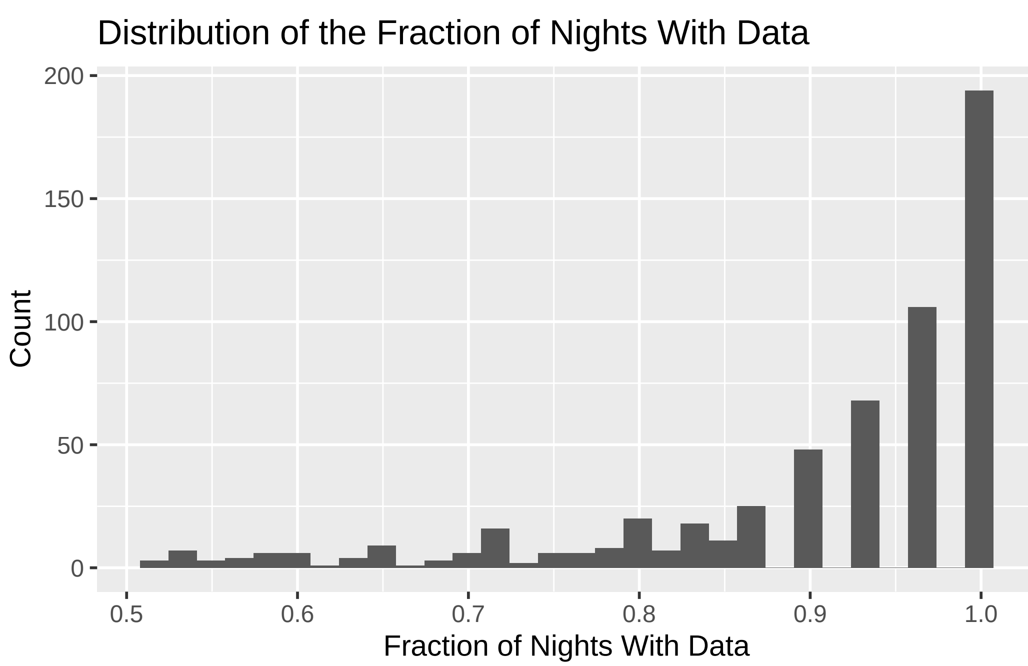
university	mean_tgpa	median_tgpa	sd_tgpa	min_tgpa	max_tgpa	count
cath_priv	3.665	3.714	0.267	2.722	4	142
public	3.401	3.500	0.518	0.350	4	249

university	mean_tgpa	median_tgpa	sd_tgpa	min_tgpa	max_tgpa	count
stem_priv	3.359	3.490	0.535	1.500	4	197

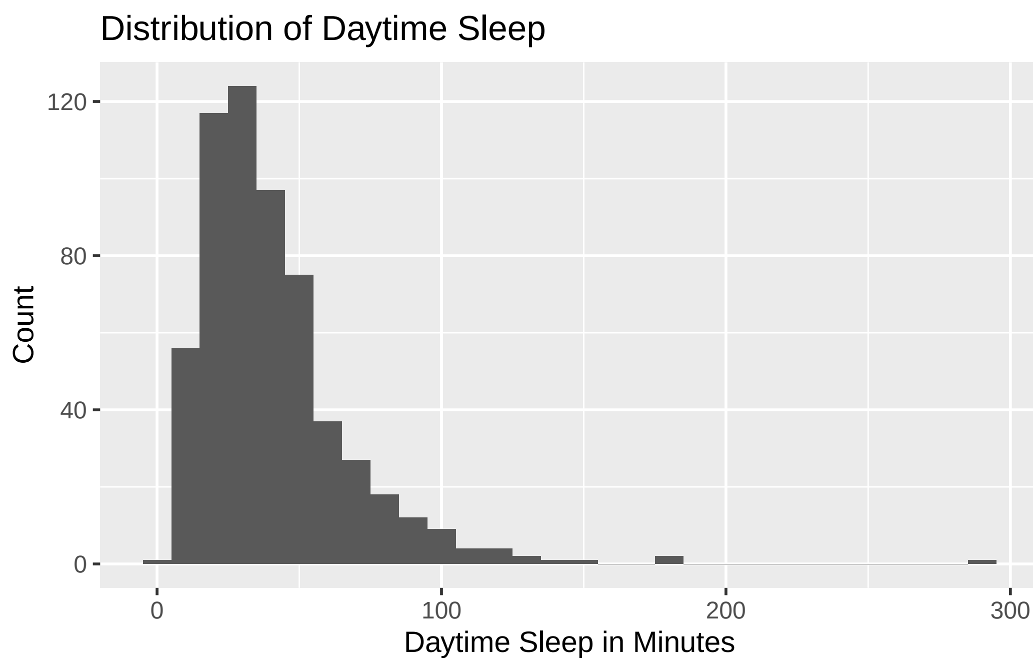
university	mean_cgpa	median_cgpa	sd_cgpa	min_cgpa	max_cgpa	count
cath_priv	3.639	3.714	0.261	2.800	4	142
public	3.429	3.501	0.400	1.588	4	249
stem_priv	3.388	3.520	0.554	1.210	4	197

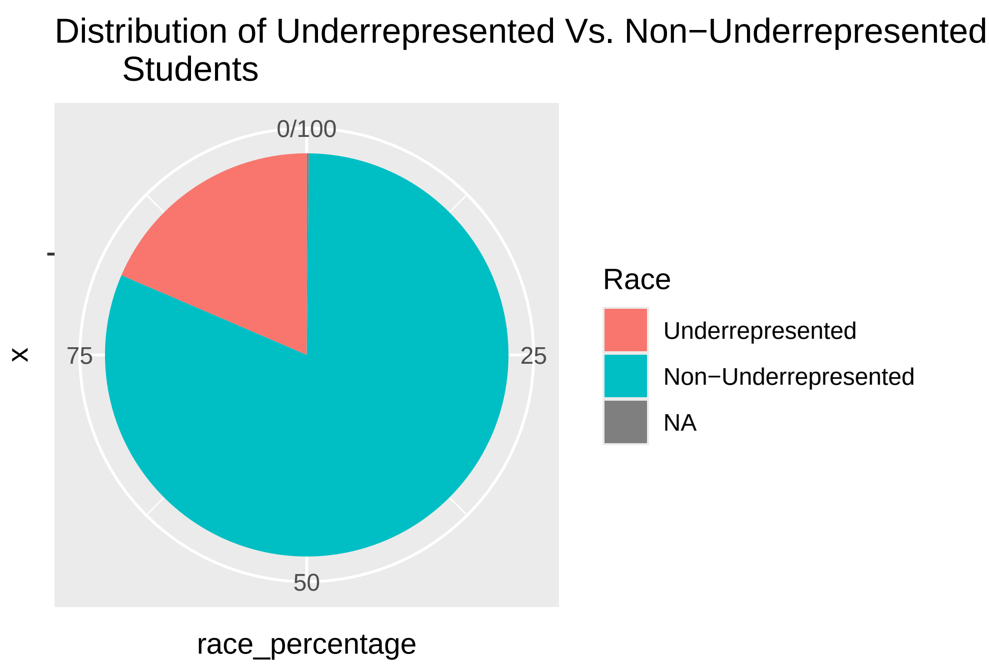


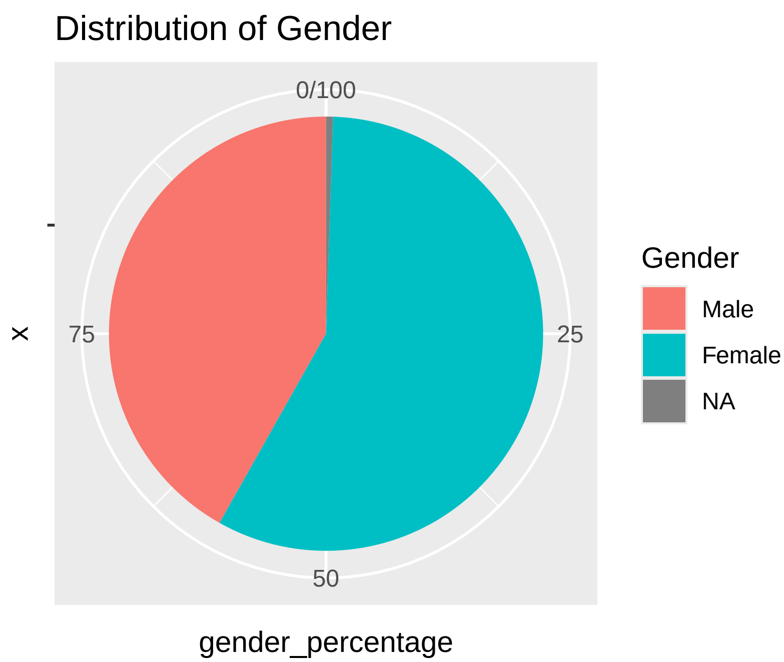
Distribution of the Bedtime Variability
Count
200
100
0
0
1
2
3
4
Mean successive squared difference of bedtime

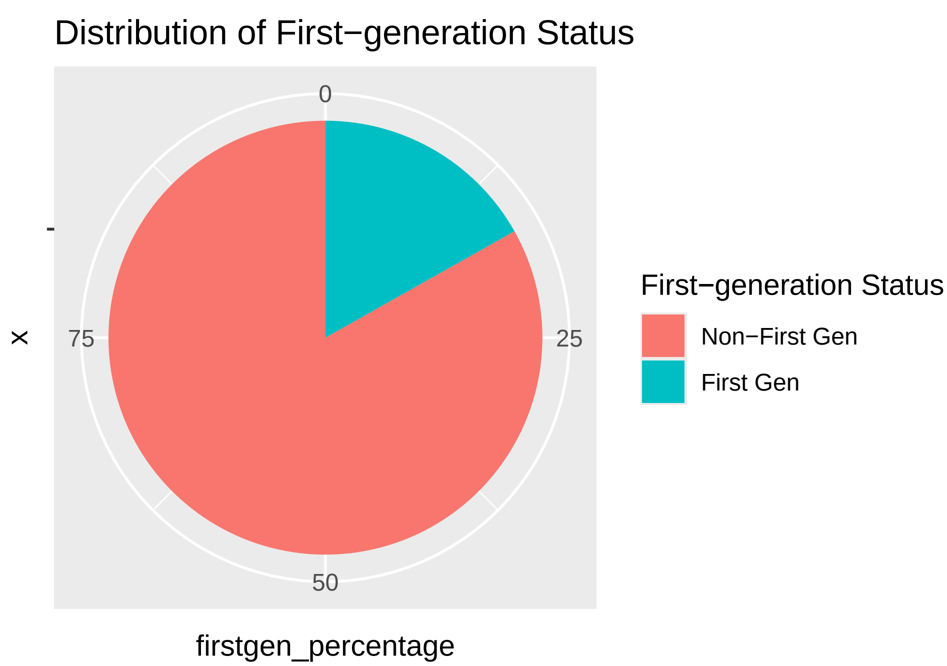


Distribution of the Fraction of Nights With Data
Count
200
150
100
50
0
0.5
0.6
0.7
0.8
0.9
1.0
Fraction of Nights With Data


Distribution of Daytime Sleep
Count
120
80
40
0
0
100
200
300
Daytime Sleep in Minutes


Distribution of Underrepresented Vs. Non−Underrepresented Students
0/100
25
50
75
x
race_percentage
Race
Underrepresented
Non−Underrepresented
NA

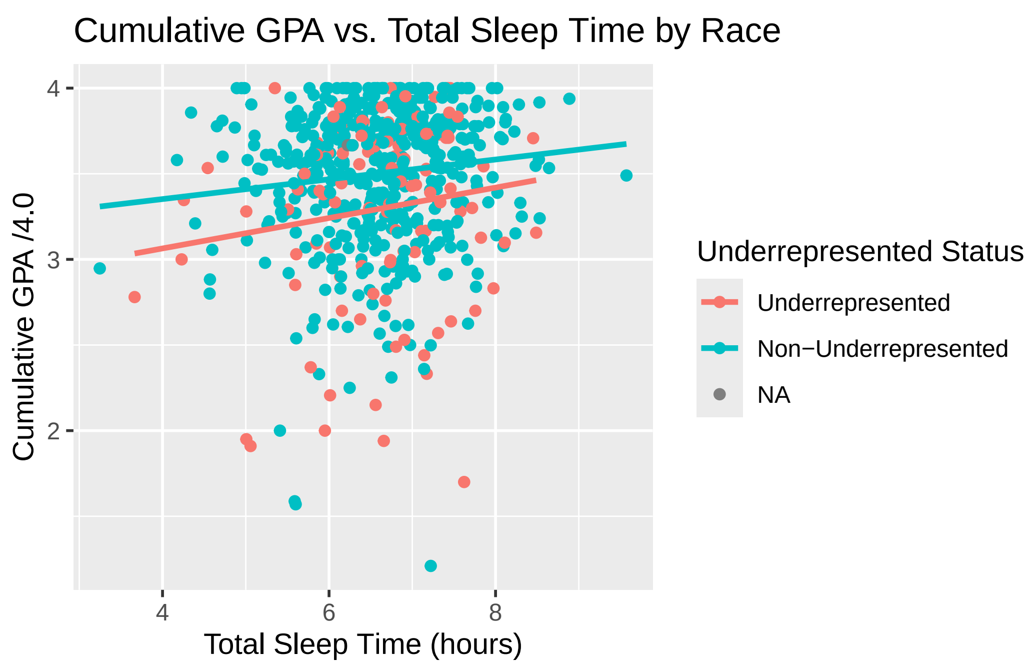


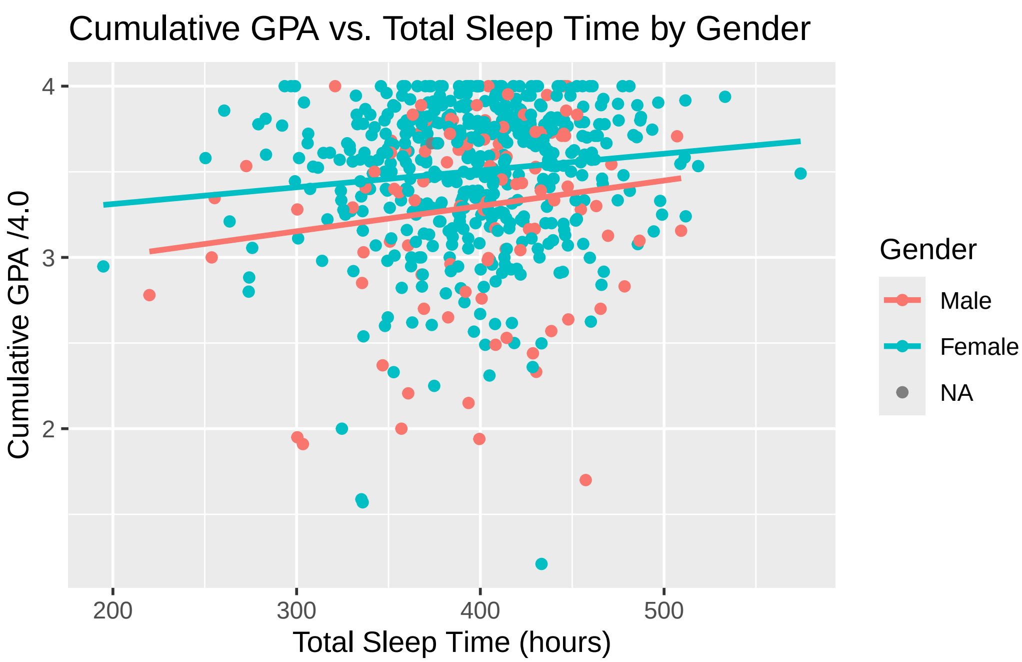


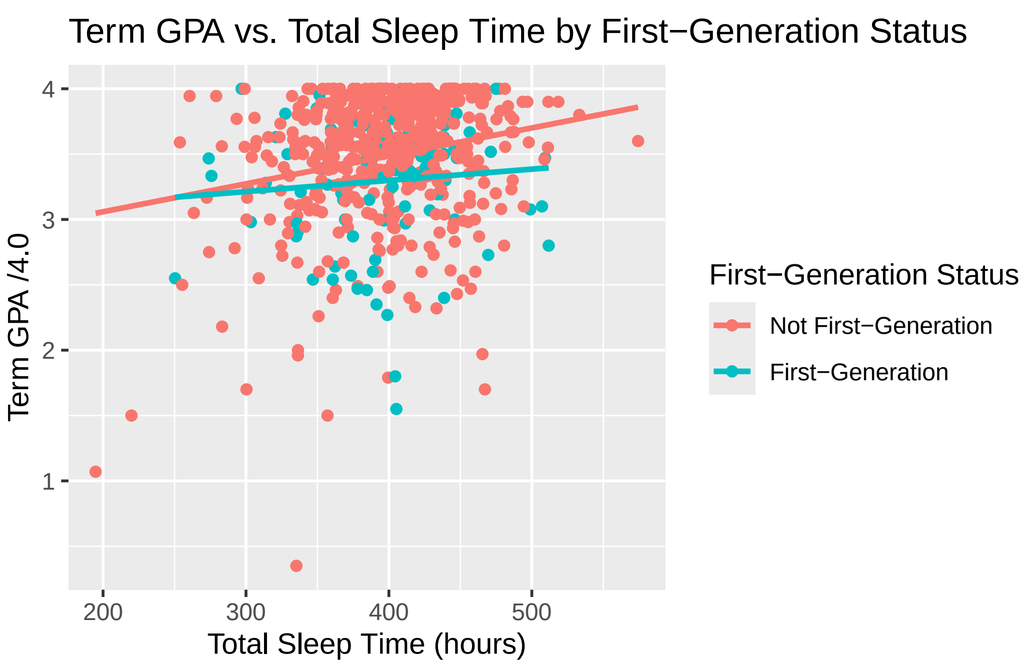
university	total_count	na_count	non_na_count
cath_priv	142	142	0
public	249	0	249

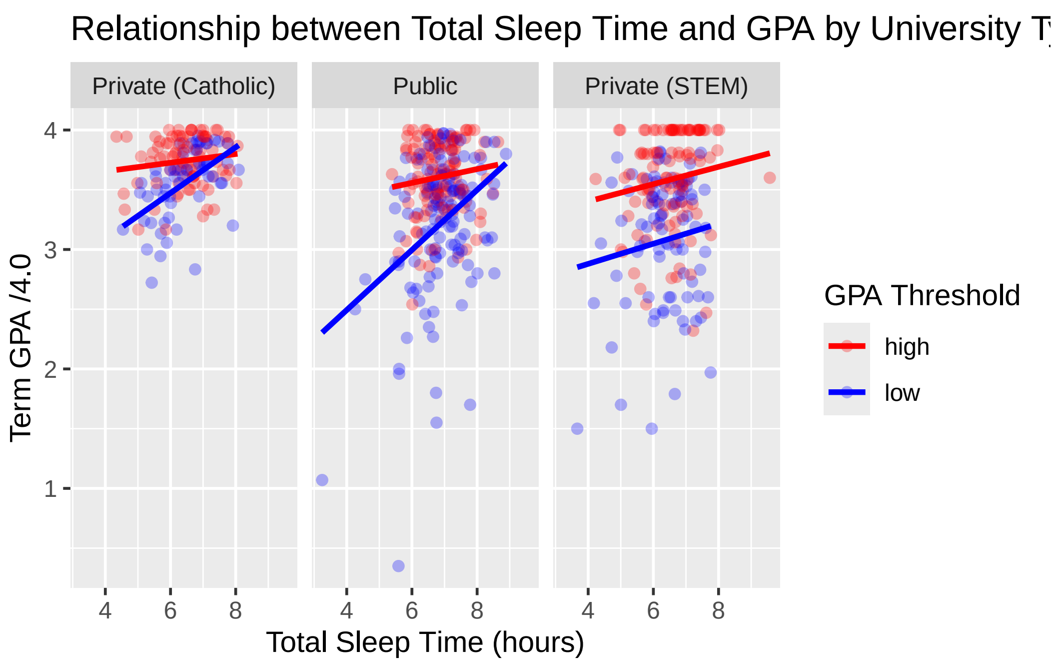
university	total_count	na_count	non_na_count
stem_priv	197	0	197

Bivariate EDA of The Response & Key Predictor Variables:

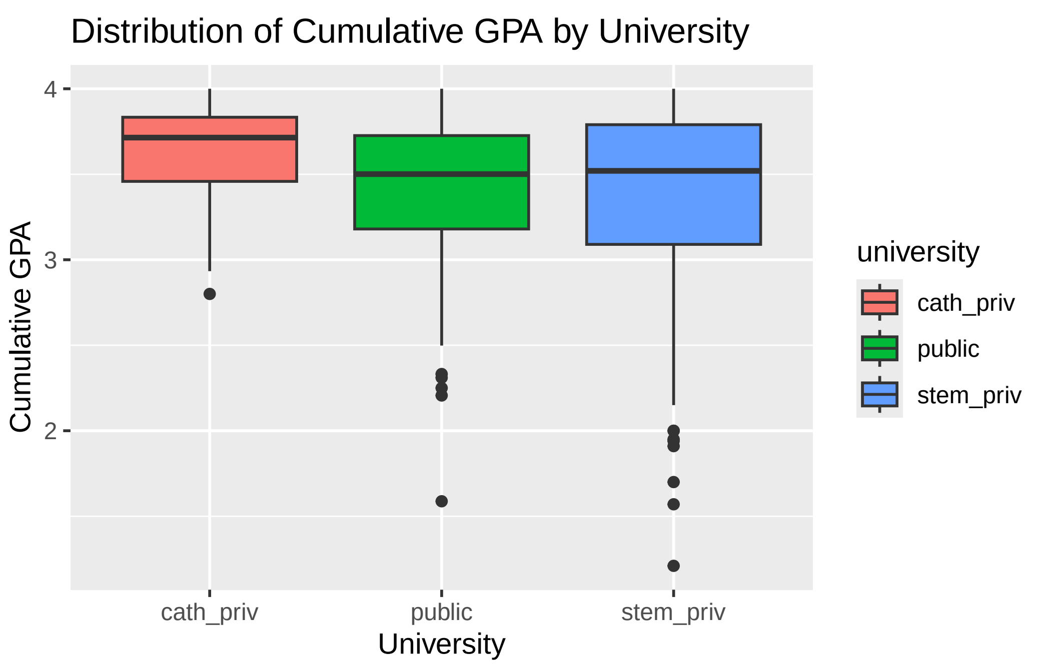



Cumulative GPA vs. Total Sleep Time by Gender
Cumulative GPA /4.0
4
3
2
200
300
400
500
Total Sleep Time (hours)
Gender
Male
Female
NA

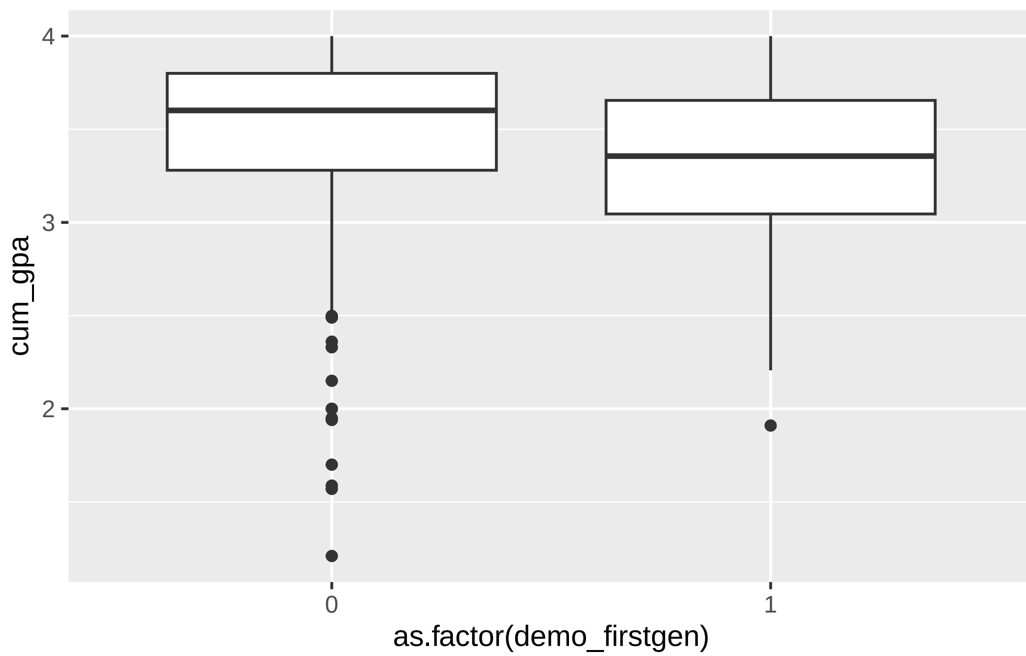

Term GPA vs. Total Sleep Time by First−Generation Status
Term GPA /4.0
4
3
2
1
200
300
400
500
Total Sleep Time (hours)
First−Generation Status
Not First−Generation
First−Generation

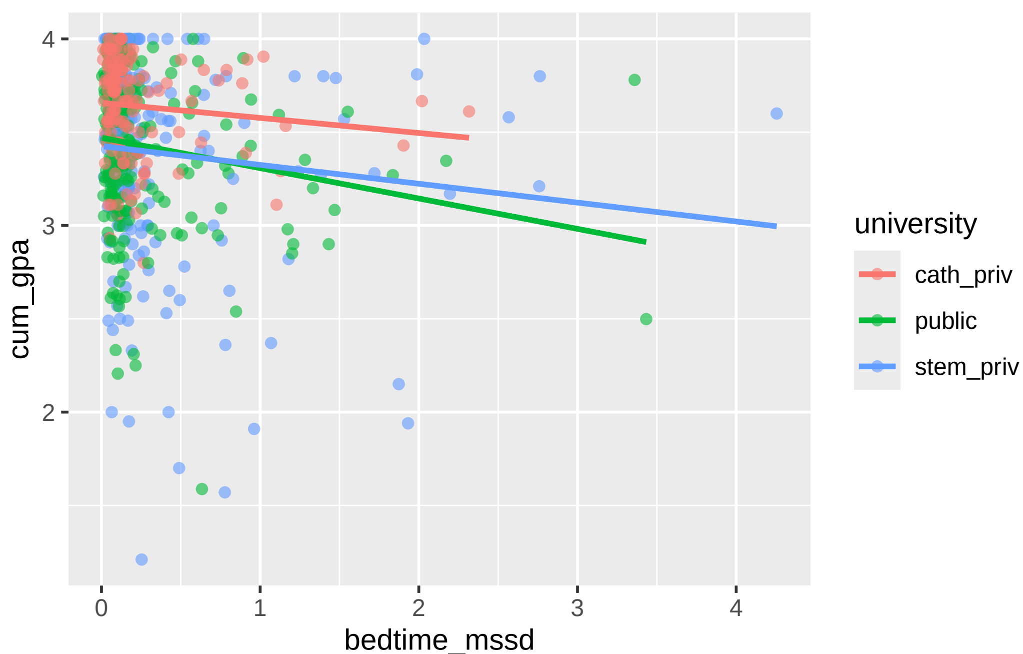


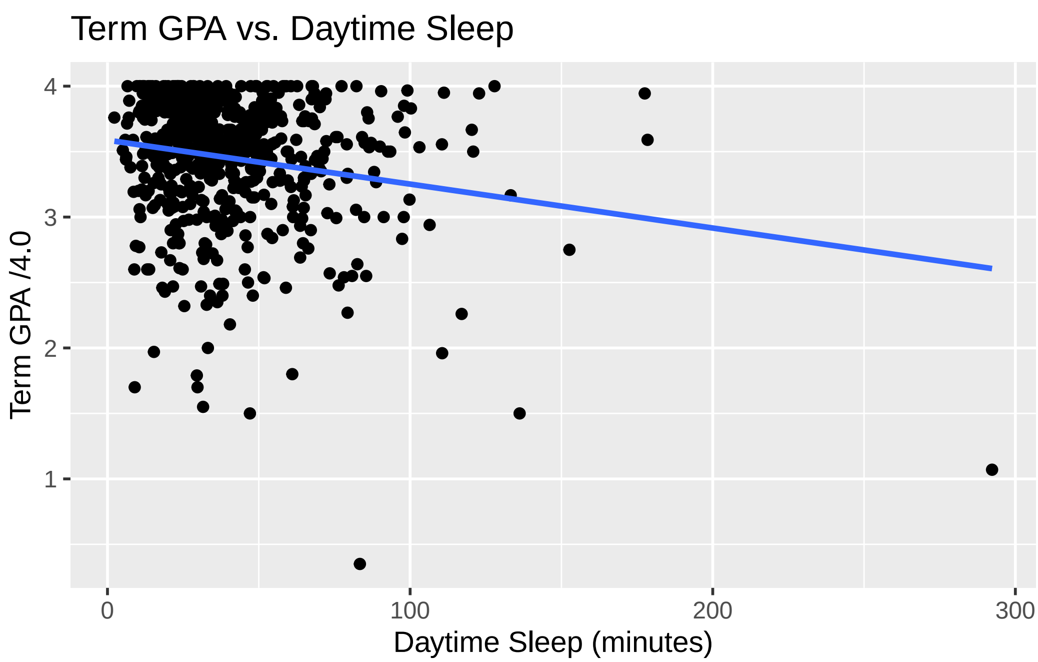
Relationship between Total Sleep Time and GPA by University T
Private (Catholic)
Public
Private (STEM)
Term GPA /4.0
Total Sleep Time (hours)
GPA Threshold
high
low

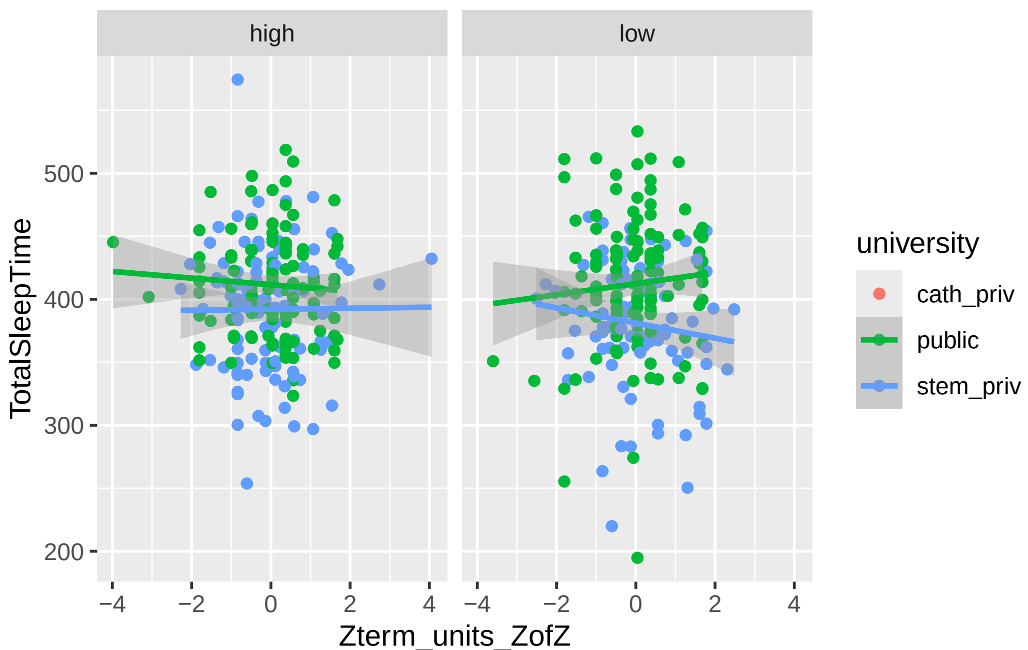


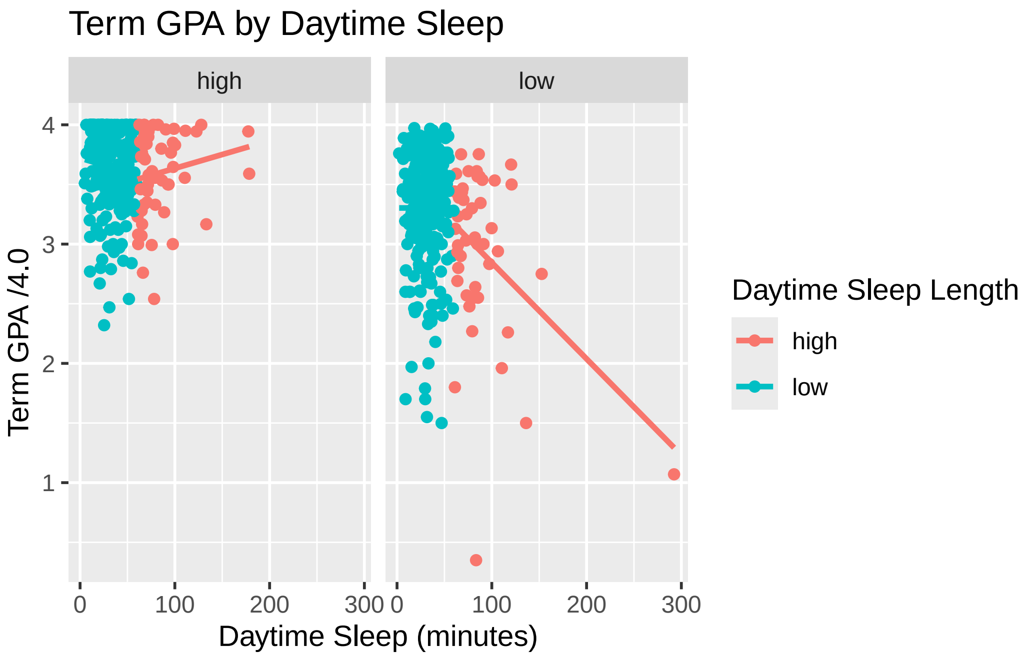
Distribution of Cumulative GPA by University
Cumulative GPA
University
cath_priv
public
stem_priv
university


4
3
2
cum_gpa
0
1
as.factor(demo_firstgen)


4
3
2
cum_gpa
0
1
2
3
4
bedtime_mssd
university
cath_priv
public
stem_priv


Term GPA vs. Daytime Sleep
Term GPA /4.0
Daytime Sleep (minutes)
0
100
200
300
1
2
3
4


high
low
TotalSleepTime
Zterm_units_ZofZ
university
cath_priv
public
stem_priv
200
300
400
500
−4
−2
0
2
4



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, term_l_cum, 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.

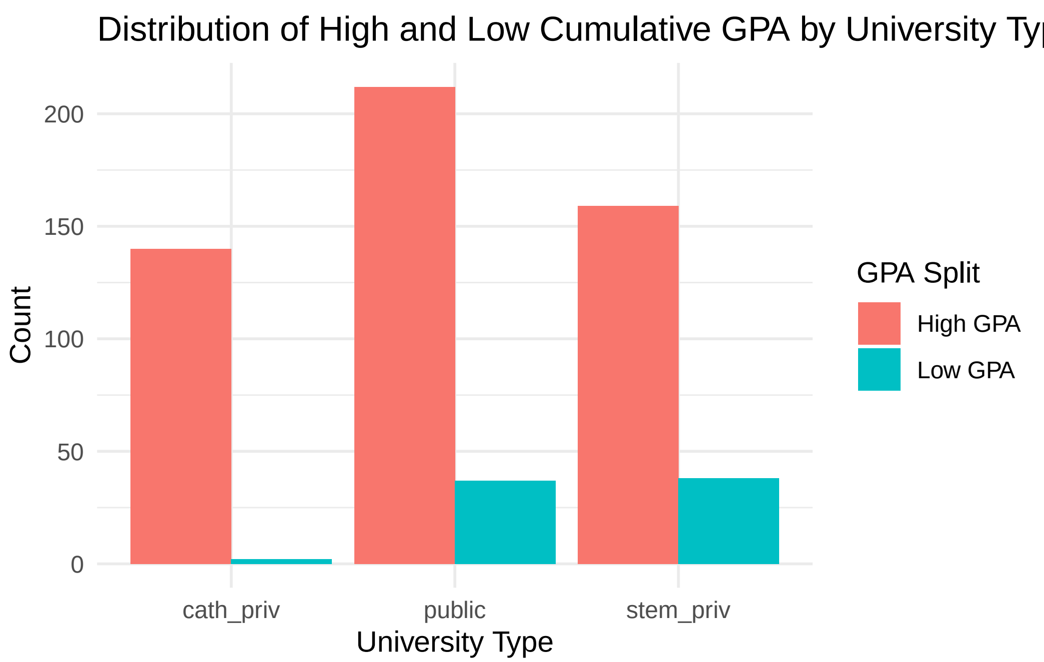
update

Regression Testing

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
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

```
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

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

        1
0.4573407

         1
0.04074673

        1
0.3756176
```

predict prob of 21.7% that

- underrepresented
- stem
- that's doing better spring sem than fall sem

has a GPA of at least 3.0.

multicolinearity check for TotalSleepTime and daytime_sleep

Questions:

- how to determine which variables to use in fitted model? trial and error or vif or other methods since we have so many
- drop in deviance test to see if high pval variables should be included in 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

```
# A tibble: 2 x 6
  term                        df.residual residual.deviance    df deviance p.value
  <chr>                             <dbl>             <dbl> <dbl>    <dbl>   <dbl>
1 as.factor(gpa_split) ~ u~           581              388.    NA    NA      NA
2 as.factor(gpa_split) ~ u~           580              386.     1     2.28   0.131
```

```
# A tibble: 2 x 6
  term                        df.residual residual.deviance    df deviance p.value
  <chr>                             <dbl>             <dbl> <dbl>    <dbl>   <dbl>
1 as.factor(gpa_split) ~ u~           581              388.    NA    NA      NA
2 as.factor(gpa_split) ~ u~           580              386.     1     2.56   0.110
```

```
# A tibble: 2 x 6
  term                        df.residual residual.deviance    df deviance p.value
  <chr>                             <dbl>             <dbl> <dbl>    <dbl>   <dbl>
1 as.factor(gpa_split) ~ u~           581              388.    NA    NA     NA
2 as.factor(gpa_split) ~ u~           580              384.     1     4.30  0.0382
```

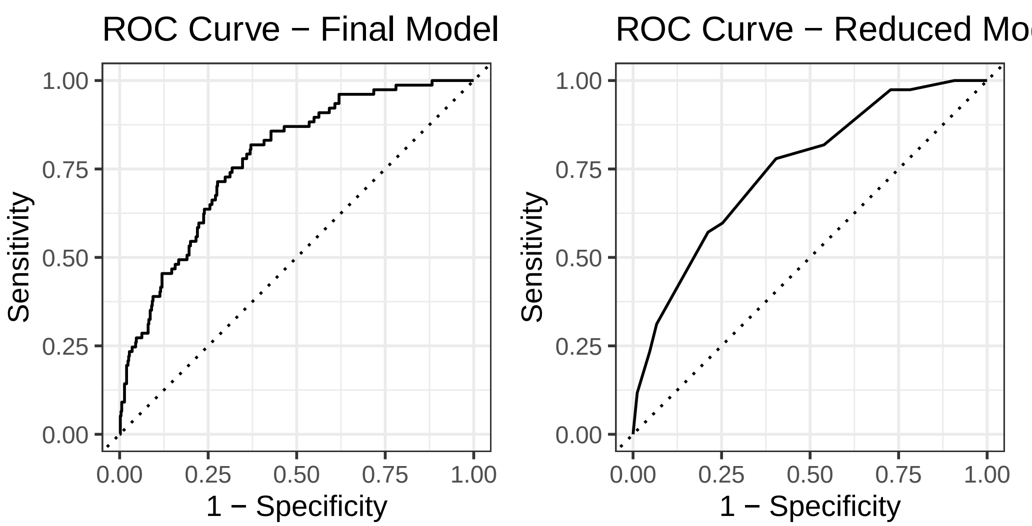
```
                   GVIF Df GVIF^(1/(2*Df))
university     1.080826  2        1.019621
demo_race      1.020008  1        1.009954
threshold_gpa  1.026319  1        1.013074
TotalSleepTime 1.079010  1        1.038754
```

```
                       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
daytime_sleep_lvl  1.103836  1        1.050636
TotalSleepTime     1.133772  1        1.064787
```

```
# 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       1.13       1                 1.06
5       1.10       1                 1.05
```

No multicollinearity issues with the final model.

ROC:




```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.778
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.751
```

One metric for determining the strength of University_final:

AIC , BIC:

```
[1] 406.2854
```

```
[1] 398.0298
```

```
[1] 428.1606
```

```
[1] 428.6549
```

Although the BIC for our final model is higher, because we are looking to predict future gpa levels from these factors, AIC is a more appropriate gauge for determining a better model. The AIC for university_final is lower than the AIC for sig_fit, therefore, we believe that university final is a better model to predict a high or low GPA, based on the factors given in the model.

Another reason why we believe that the university final model is better than the reduced model is because the ROC curve fits the final model better than the reduced model. The area under the curve for the final model is 77.8%, whereas for the reduced model, it is 75.1%. This tells us that the final model maximizes sensitivity and minimizes 1 - specificity better than the reduced model.

Things that still need to be done:

Explain the drop in deviance test results

Explain transformation of variables:

- daytime sleep lvl
- gpa_threshold
- gpa_split
- university (check this explanation)

Explain why we chose gpa_split as our response variable

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

Results Section:

In this section, you will output the final model and include a brief discussion of the model assumptions, diagnostics, and any relevant model fit statistics.

This section also includes initial interpretations and conclusions drawn from the model.

Introduction and data:

methodology:

results: