Your project title

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.

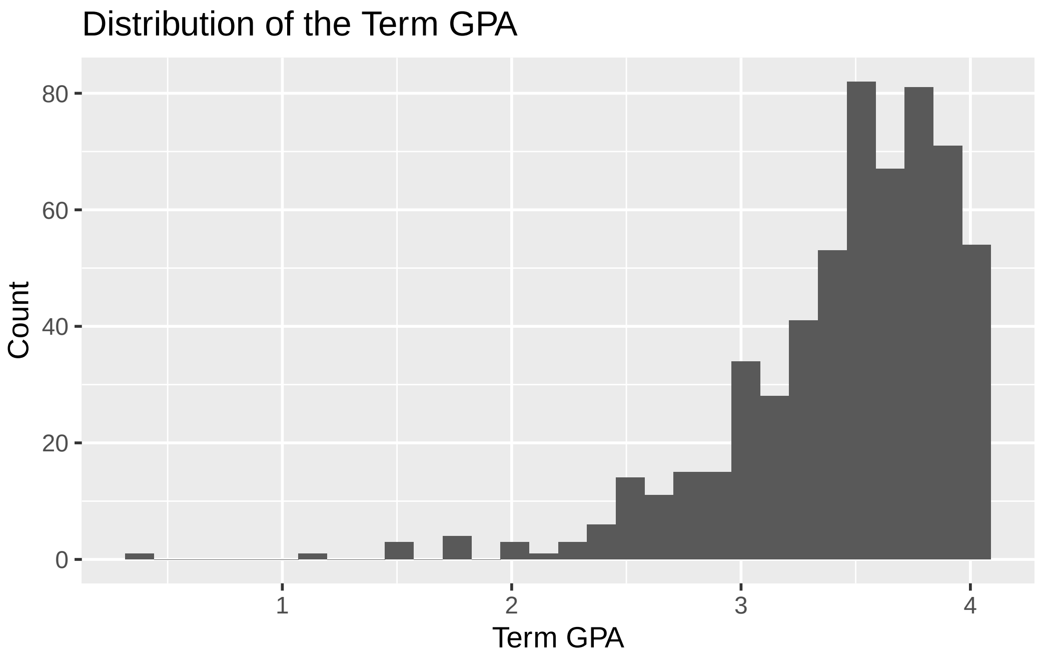
There are 634 observations, representing the 634 participants in this study. Race is a binary variable separated into underrepresented students and non-underrepresented students with 0 being underrepresented and 1 being non-underrepresented. Students are considered underrepresented if either parent is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

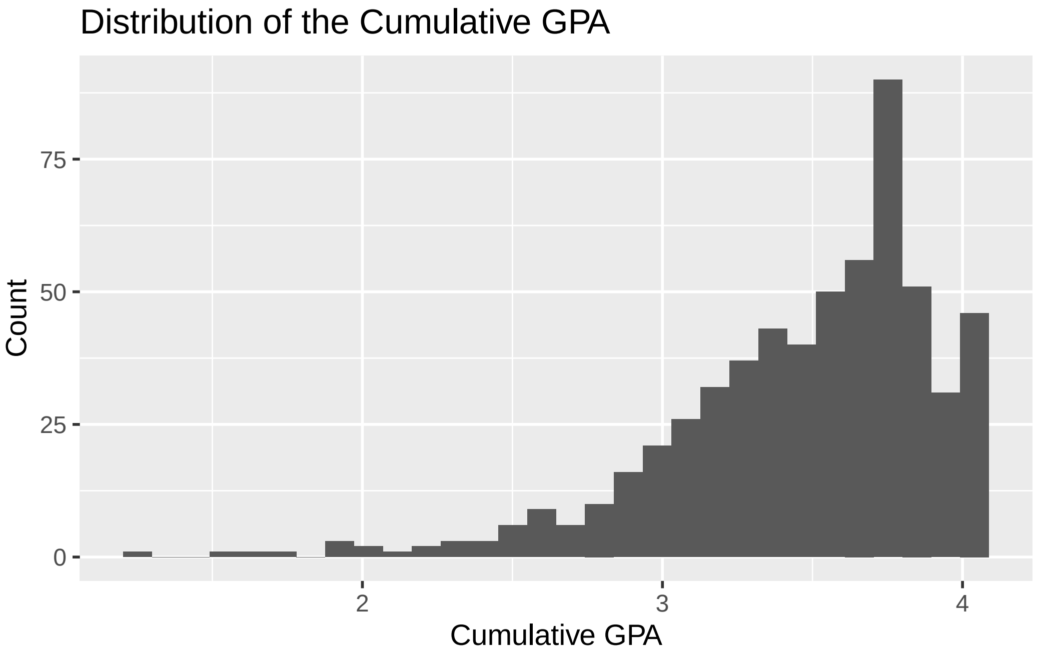
Clean Data

```
Rows: 588
Columns: 16
```

```
$ subject_id            <dbl> 185, 158, 209, 102, 174, 184, 255, 265, 343, 137~
$ study                 <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
$ cohort                <chr> "lac1", "lac1", "lac1", "lac1", "lac1", "lac1", ~
$ demo_race             <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ demo_gender           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, ~
$ demo_firstgen         <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ~
$ bedtime_mssd          <dbl> 0.11672695, 0.14168084, 1.52928949, 0.13014845, ~
$ TotalSleepTime        <dbl> 432.2000, 391.9310, 344.3043, 392.6207, 423.4211~
$ midpoint_sleep        <dbl> 458.6600, 364.4655, 560.8913, 416.4828, 368.7632~
$ frac_nights_with_data <dbl> 0.8620690, 1.0000000, 0.7931034, 1.0000000, 0.65~
$ daytime_sleep         <dbl> 24.160000, 13.137931, 14.956522, 54.551724, 10.5~
$ cum_gpa               <dbl> 3.00, 3.66, 3.57, 3.61, 3.21, 3.20, 3.40, 3.86, ~
$ term_gpa              <dbl> 3.38, 2.60, 3.07, 3.56, 4.00, 3.36, 3.19, 3.28, ~
$ term_units            <dbl> 73, 64, 63, 61, 61, 60, 60, 60, 60, 59, 59, 58, ~
$ Zterm_units_ZofZ      <dbl> 4.0552949, 2.4825341, 2.3077829, 1.9582805, 1.95~
$ university            <chr> "stem_priv", "stem_priv", "stem_priv", "stem_pri~
```

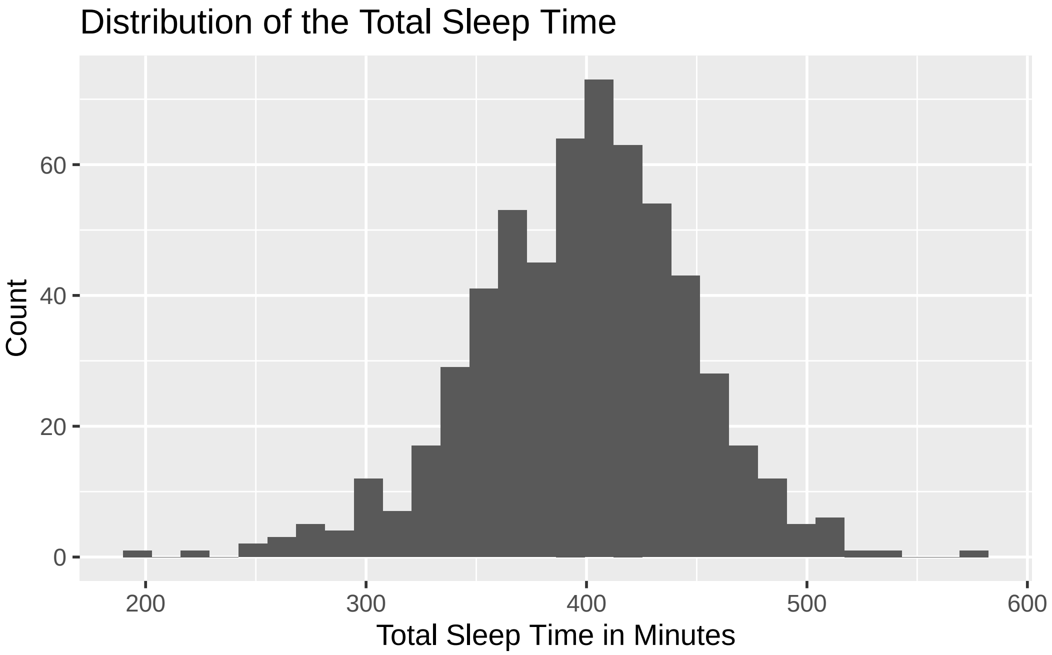
Univariate EDA of The Response & Key Predictor Variables

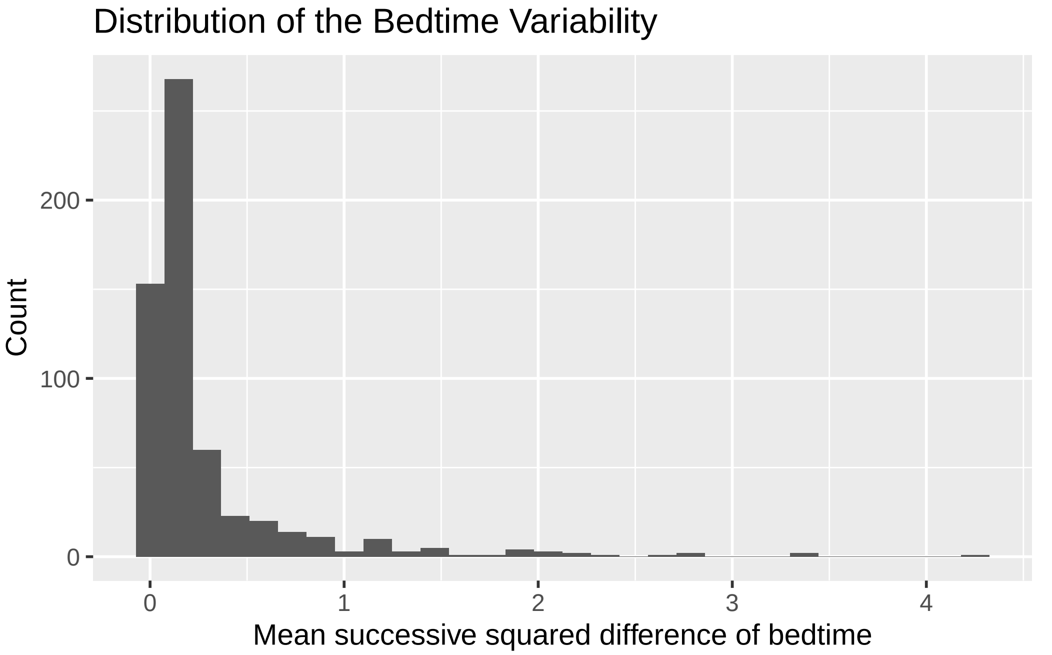


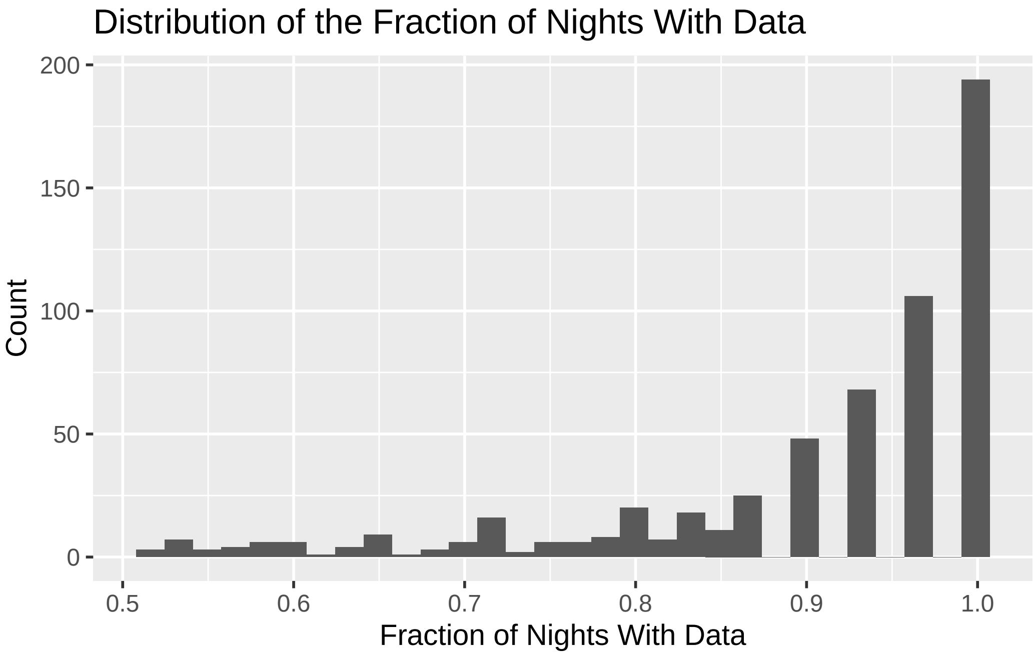


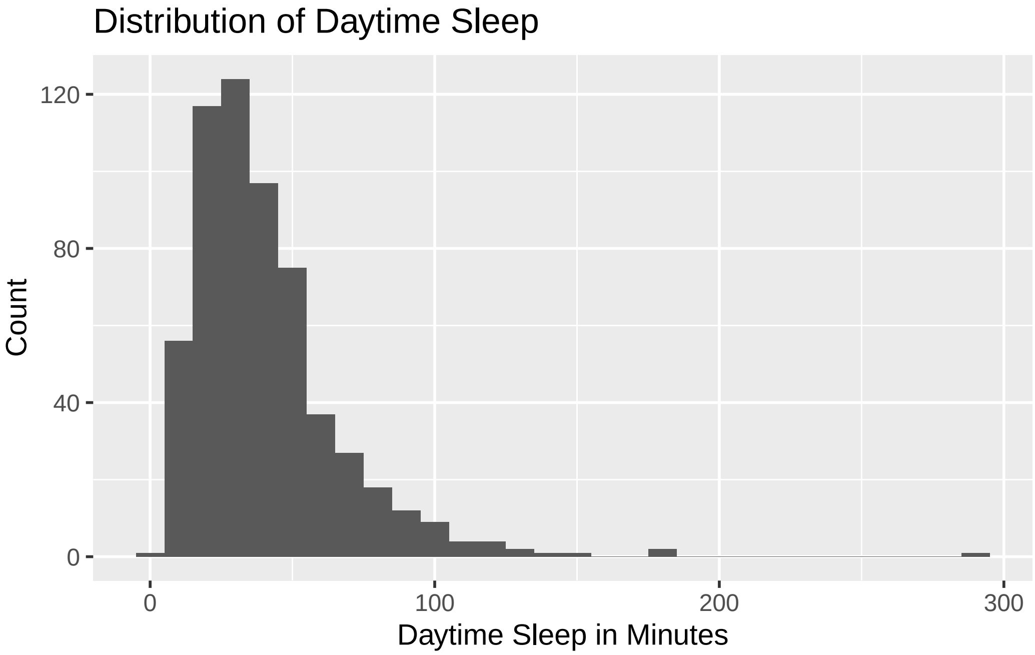
```
# A tibble: 3 x 7
  university mean_tgpa median_tgpa sd_tgpa min_tgpa max_tgpa count
  <chr>          <dbl>       <dbl>   <dbl>    <dbl>    <dbl> <int>
1 cath_priv       3.66        3.71   0.267     2.72        4   142
2 public          3.40        3.5    0.518     0.35        4   249
3 stem_priv       3.36        3.49   0.535     1.5         4   197
```

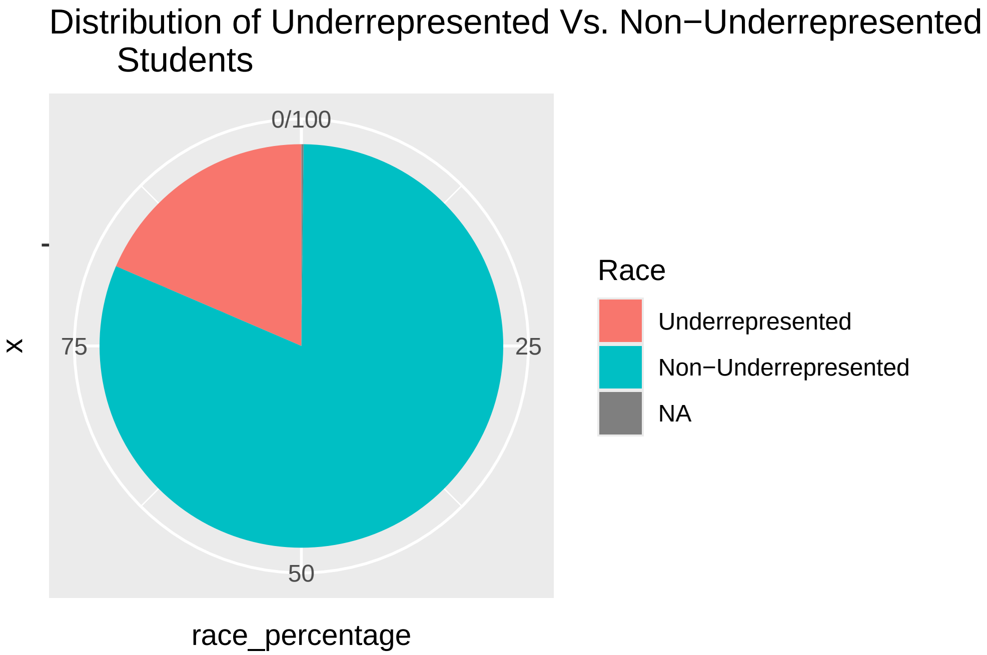
```
# A tibble: 3 x 7
  university mean_cgpa median_cgpa sd_cgpa min_cgpa max_cgpa count
  <chr>          <dbl>       <dbl>   <dbl>    <dbl>    <dbl> <int>
1 cath_priv       3.64        3.71   0.261     2.80        4   142
2 public          3.43        3.50   0.400     1.59        4   249
3 stem_priv       3.39        3.52   0.554     1.21        4   197
```

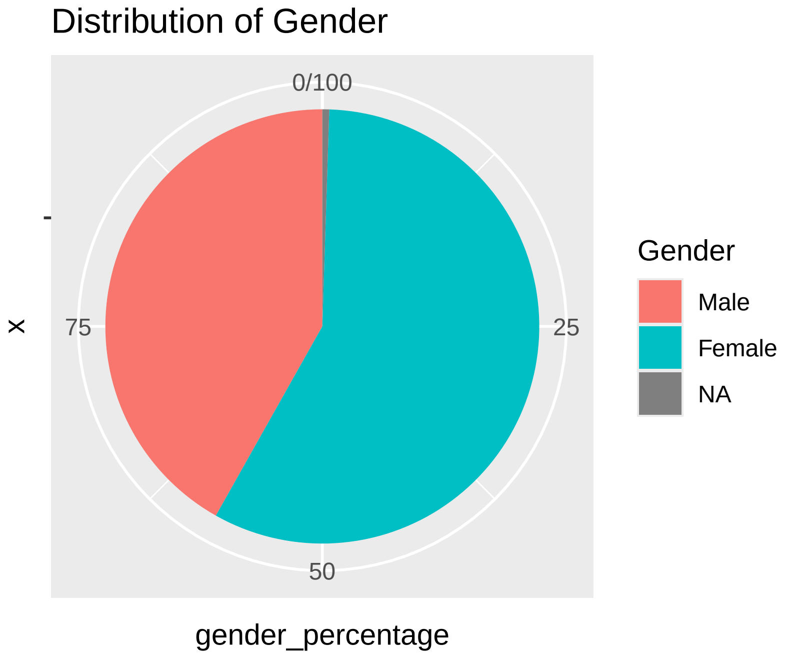

Distribution of the Total Sleep Time
Count
0
20
40
60
200
300
400
500
600
Total Sleep Time in Minutes

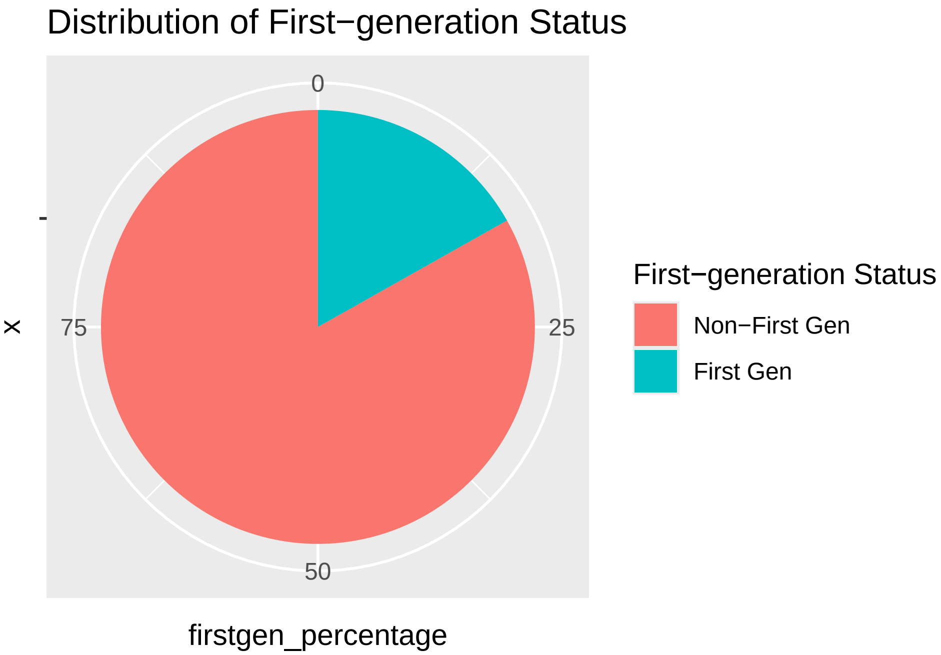

Distribution of the Bedtime Variability
Count
0
100
200
0
1
2
3
4
Mean successive squared difference of bedtime


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


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

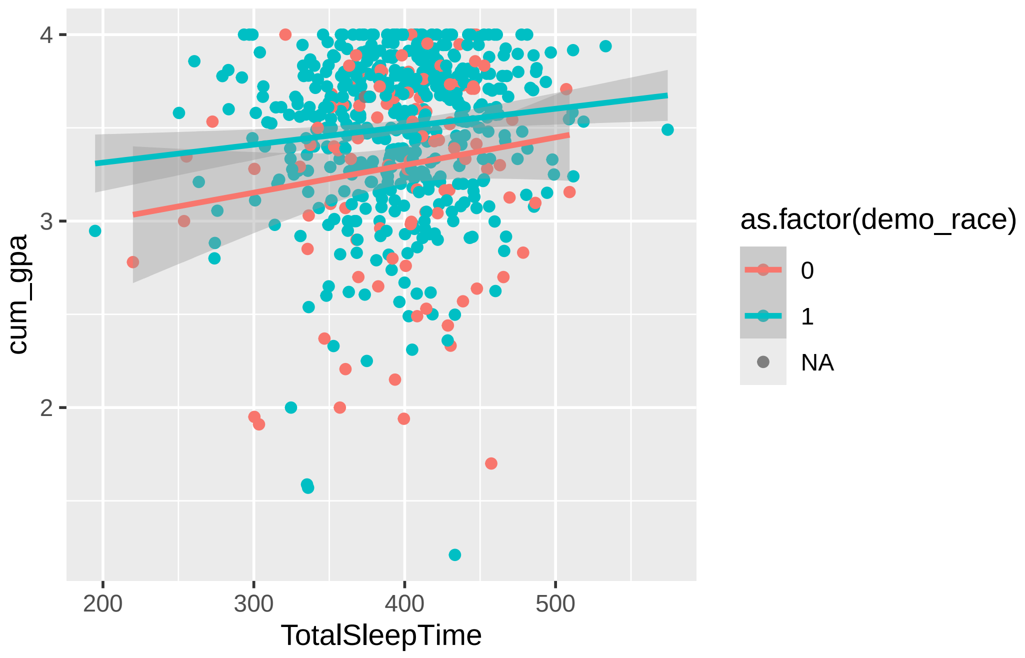

Distribution of Underrepresented Vs. Non−Underrepresented Students
0/100
25
50
75
x
race_percentage
Race
Underrepresented
Non−Underrepresented
NA

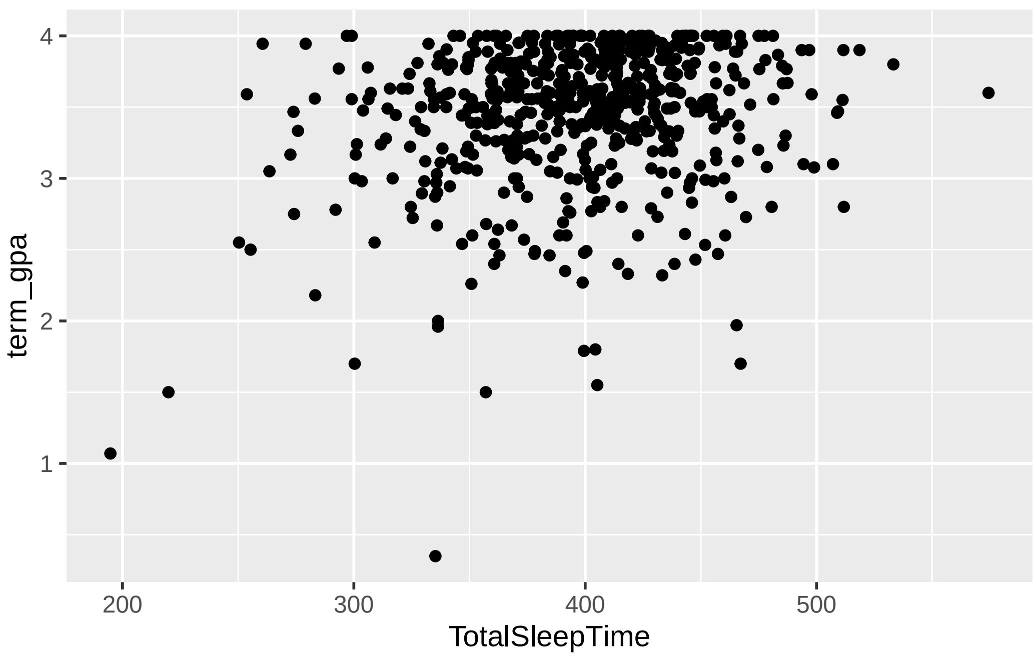

Distribution of Gender
0/100
25
50
75
x
gender_percentage
Gender
Male
Female
NA

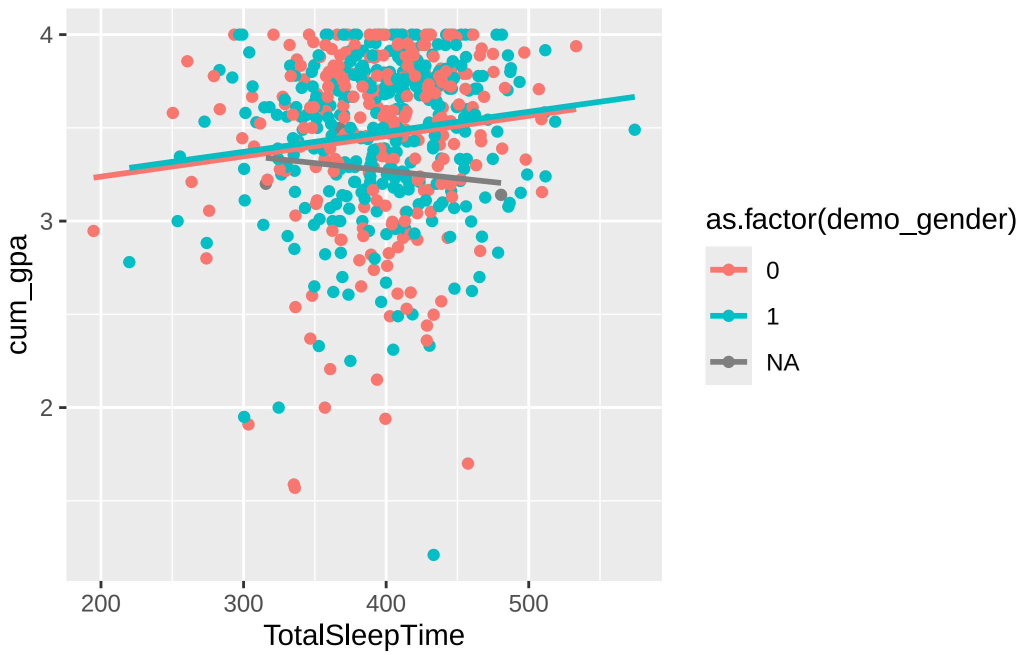


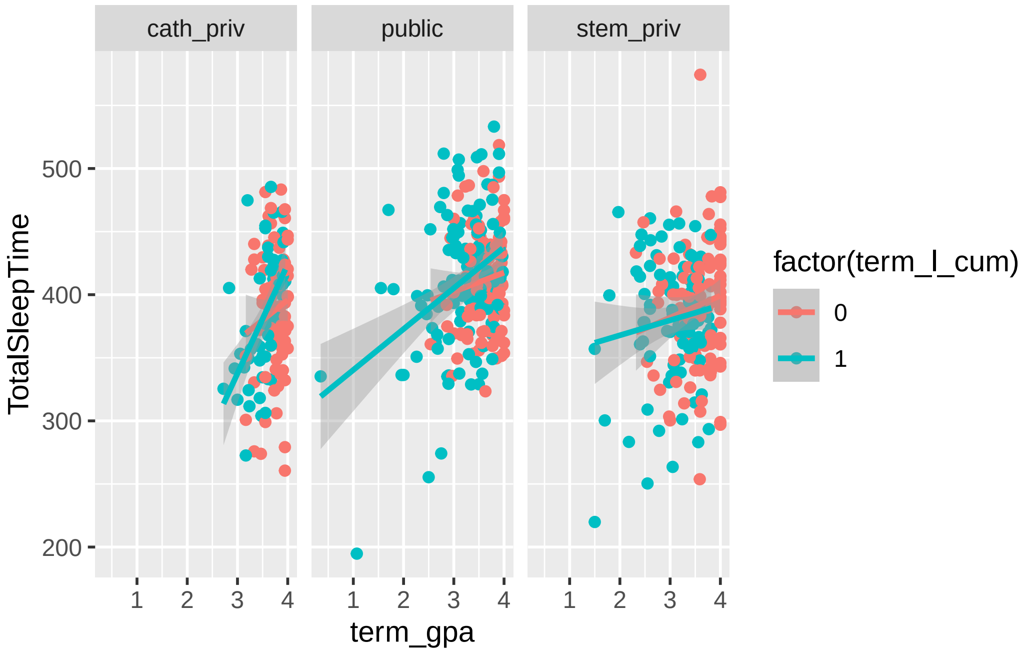
```
# A tibble: 3 x 4
  university total_count na_count non_na_count
  <chr>            <int>    <int>        <int>
1 cath_priv          142      142            0
2 public             249        0          249
3 stem_priv          197        0          197
```

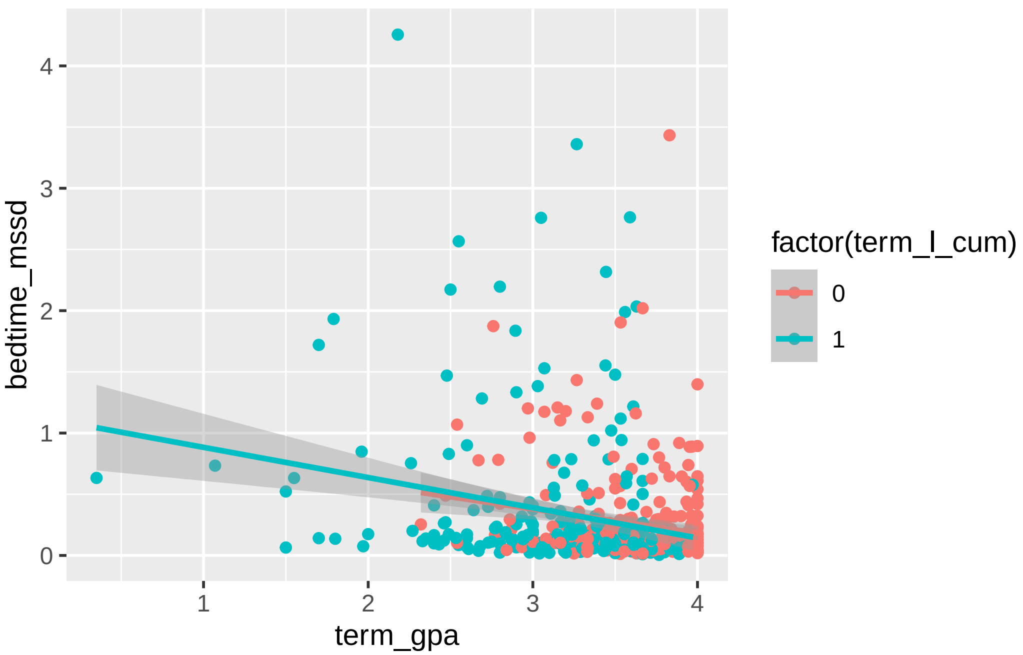
Bivariate EDA of The Response & Key Predictor Variables

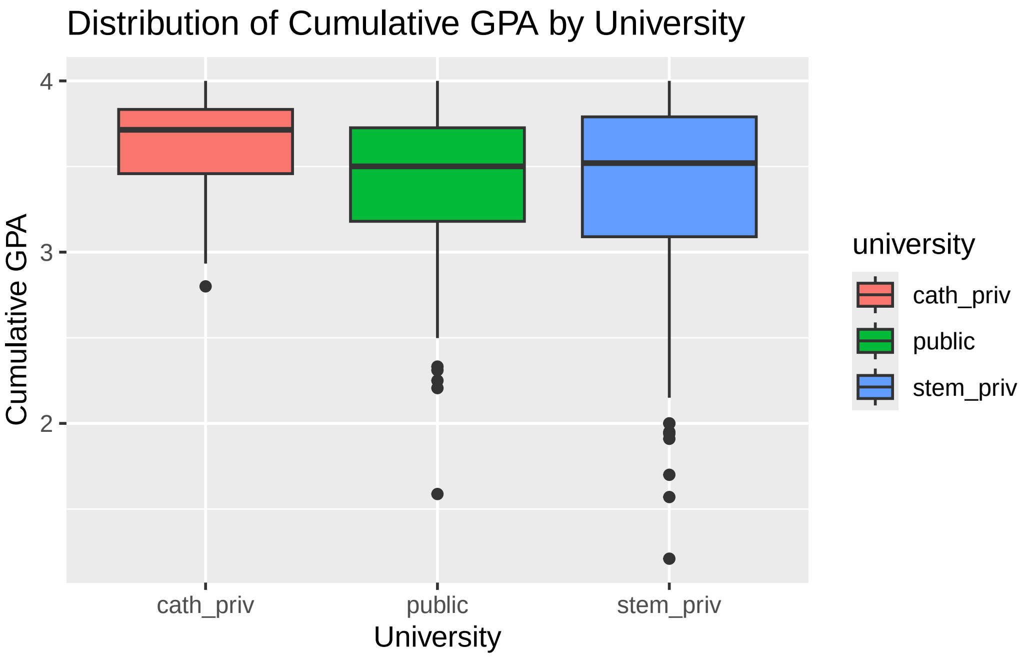

4
3
2
cum_gpa
200
300
400
500
TotalSleepTime
as.factor(demo_race)
0
1
NA

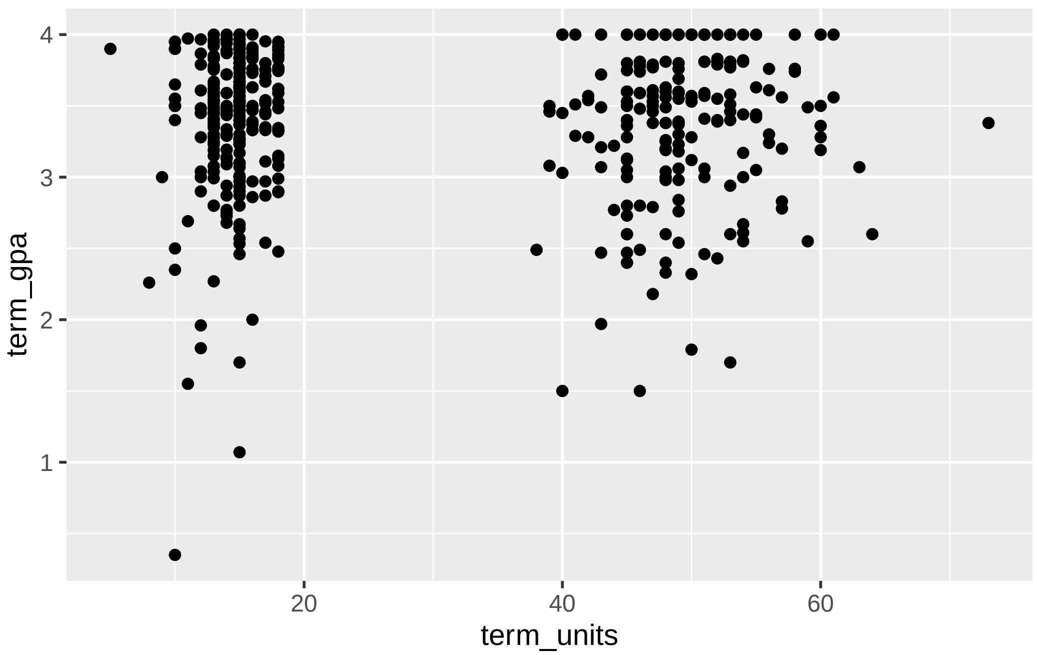

4
3
2
1
term_gpa
200
300
400
500
TotalSleepTime

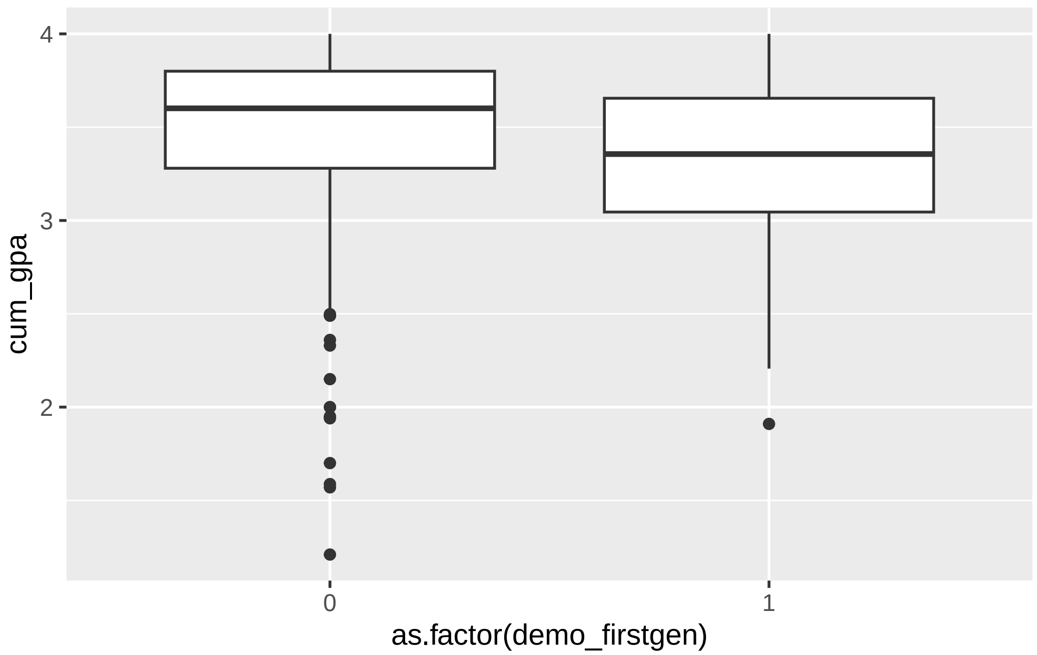

cum_gpa
4
3
2
200
300
400
500
TotalSleepTime
as.factor(demo_gender)
0
1
NA

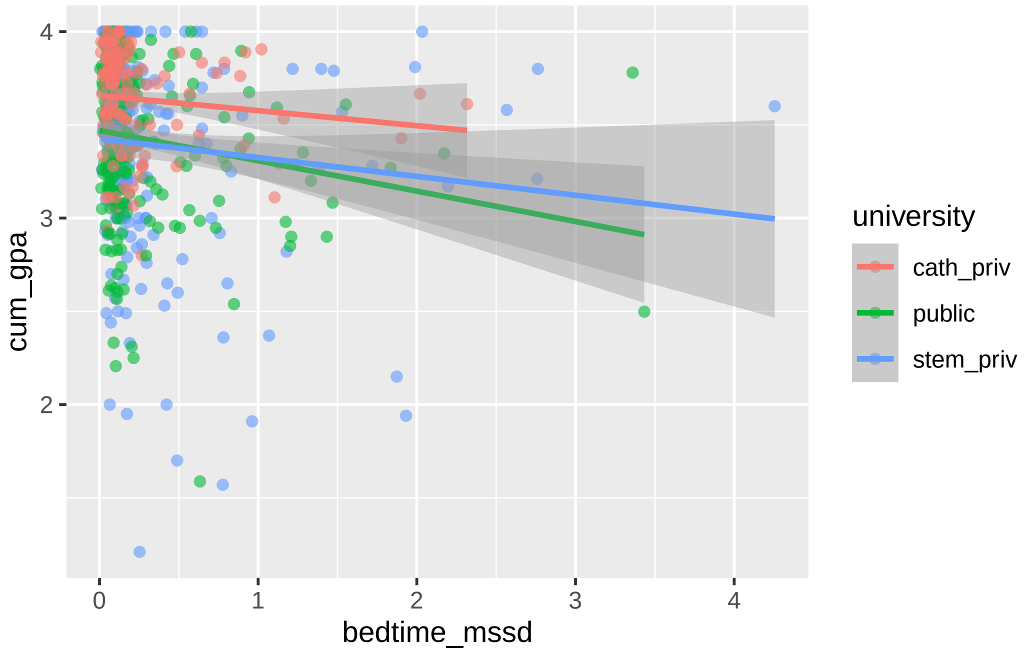

cath_priv
public
stem_priv
TotalSleepTime
500
400
300
200
1
2
3
4
term_gpa
factor(term_l_cum)
0
1

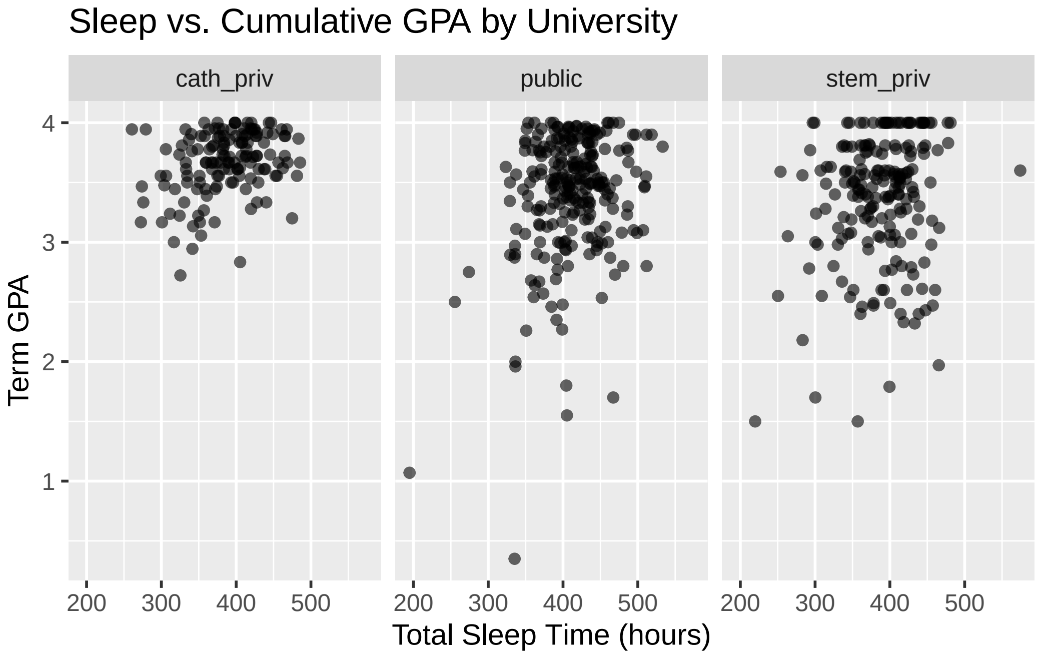

bedtime_mssd
0
1
2
3
4
term_gpa
1
2
3
4
factor(term_l_cum)
0
1

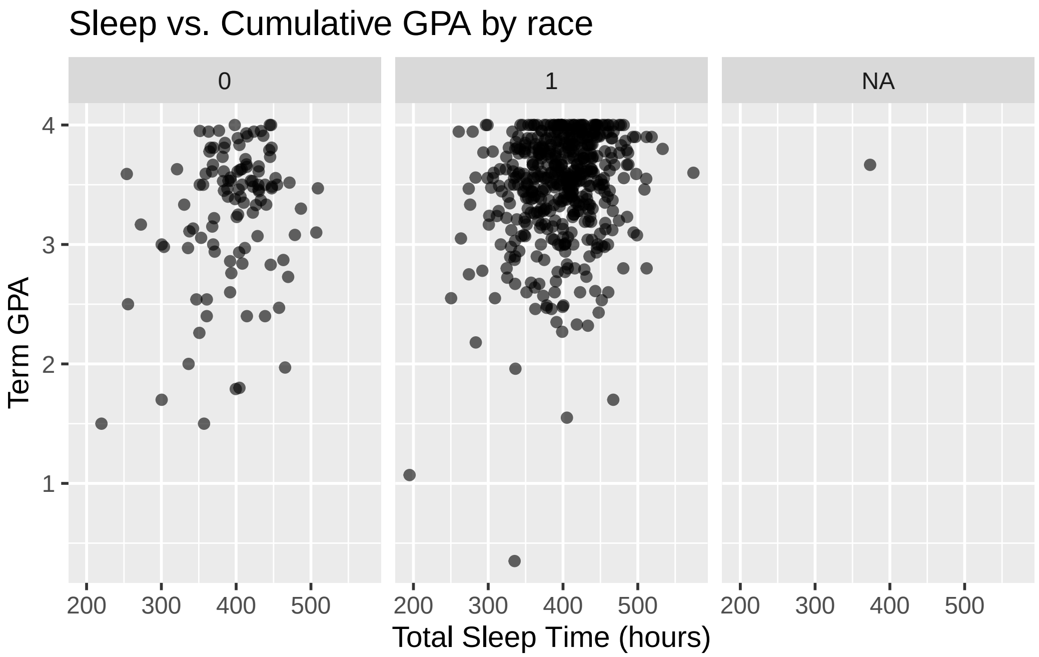

Distribution of Cumulative GPA by University
Cumulative GPA
2
3
4
cath_priv
public
stem_priv
University
university
cath_priv
public
stem_priv

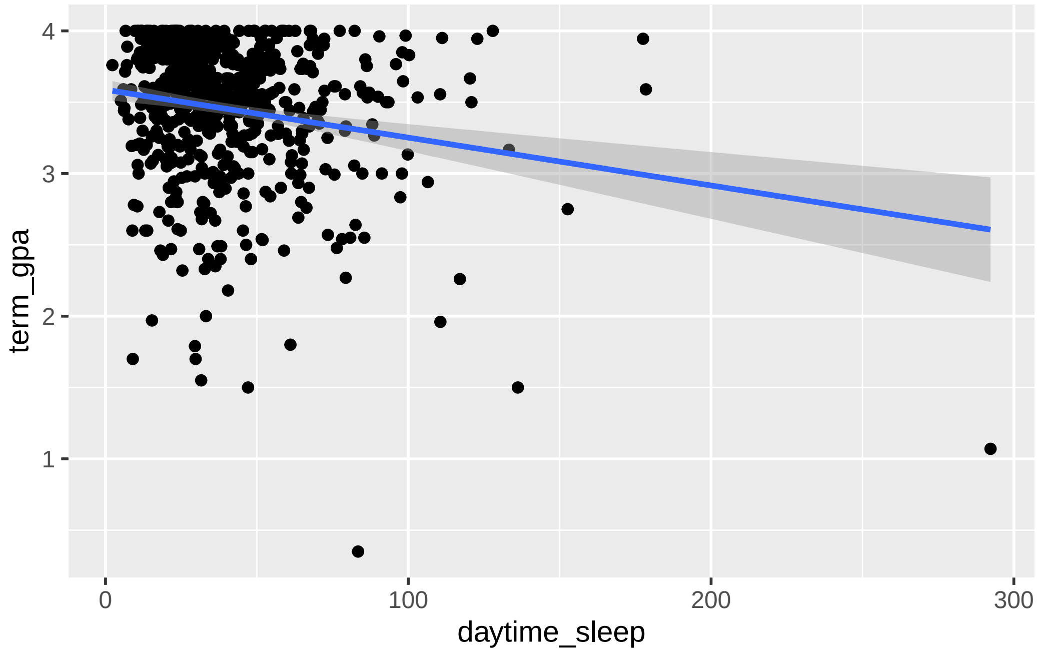

term_gpa
4
3
2
1
20
40
60
term_units


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


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


Sleep vs. Cumulative GPA by University
cath_priv
public
stem_priv
Term GPA
4
3
2
1
200
300
400
500
200
300
400
500
200
300
400
500
Total Sleep Time (hours)





```
# A tibble: 588 x 16
   subject_id study cohort demo_race demo_gender demo_firstgen bedtime_mssd
        <dbl> <dbl> <chr>      <dbl>       <dbl>         <dbl>        <dbl>
```

```
 1        185     5 lac1           1           1             0       0.117
 2        158     5 lac1           0           1             0       0.142
 3        209     5 lac1           1           1             0       1.53
 4        102     5 lac1           0           1             1       0.130
 5        174     5 lac1           1           1             0       0.130
 6        184     5 lac1           1           1             0       0.209
 7        255     5 lac1           1           1             0       0.675
 8        265     5 lac1           1           1             0       0.130
 9        343     5 lac1           1           0             0       1.48
10        137     5 lac1           1           1             0       0.0850
# i 578 more rows
# i 9 more variables: TotalSleepTime <dbl>, midpoint_sleep <dbl>,
#   frac_nights_with_data <dbl>, daytime_sleep <dbl>, cum_gpa <dbl>,
#   term_gpa <dbl>, term_units <dbl>, Zterm_units_ZofZ <dbl>, university <chr>
```

notes: stem_priv cluster is entirely lower than public cluster