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

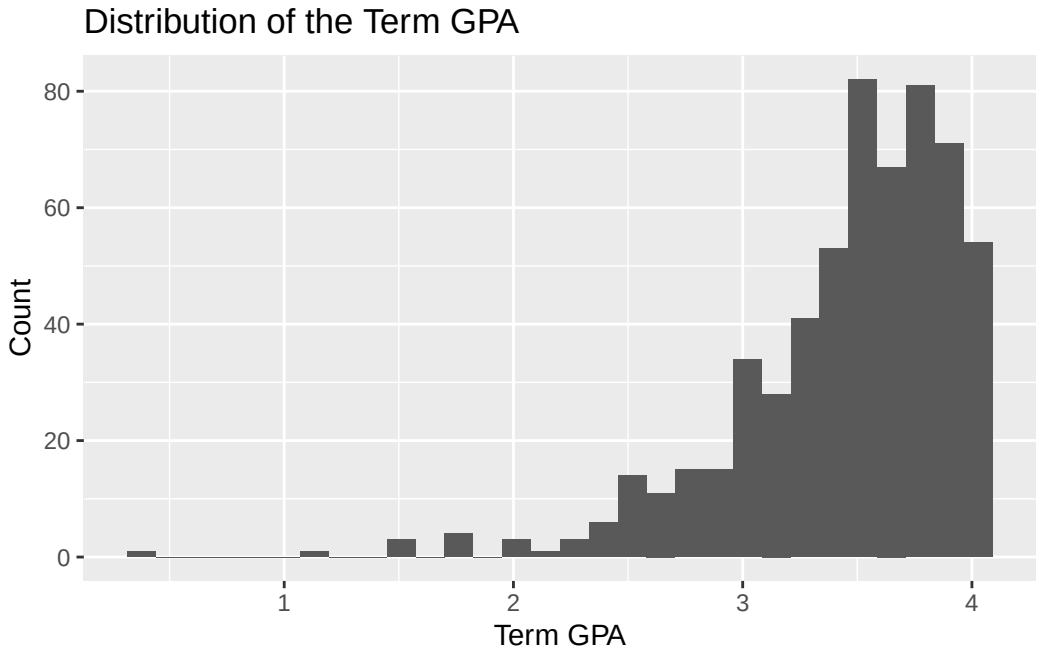
Description of the data set and key variables.

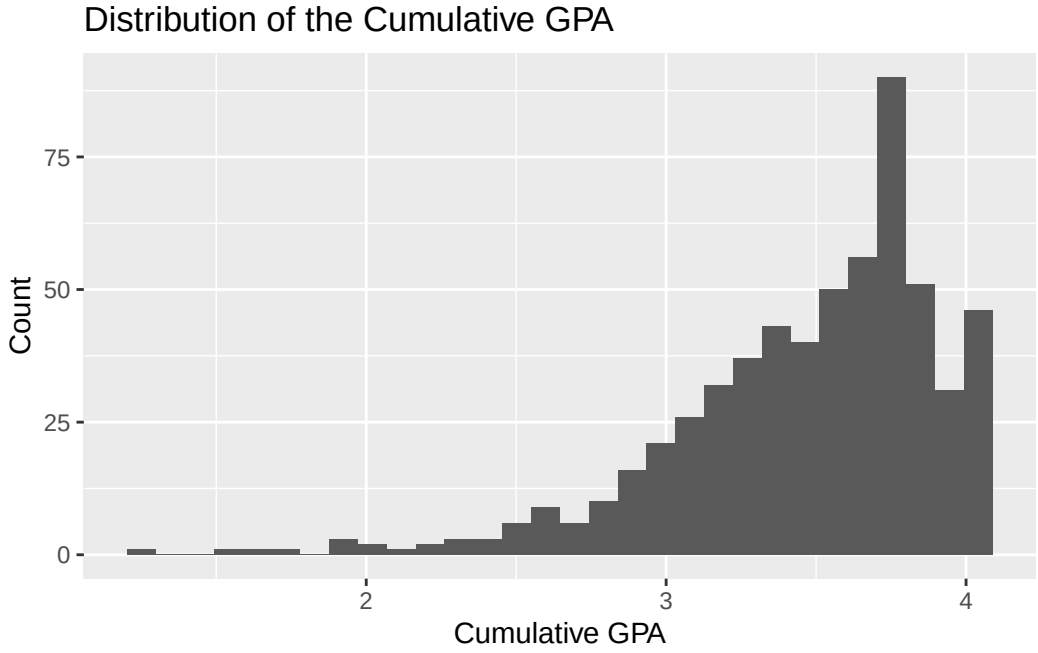
The data was originally collected in 2019, with the participants being first-year students at the following three universities: Carnegie Mellon University (CMU), a STEM-focused private university, The University of Washington (UW), a large public university, and Notre Dame University (ND), a private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

Univariate EDA of The Response & Key Predictor Variables

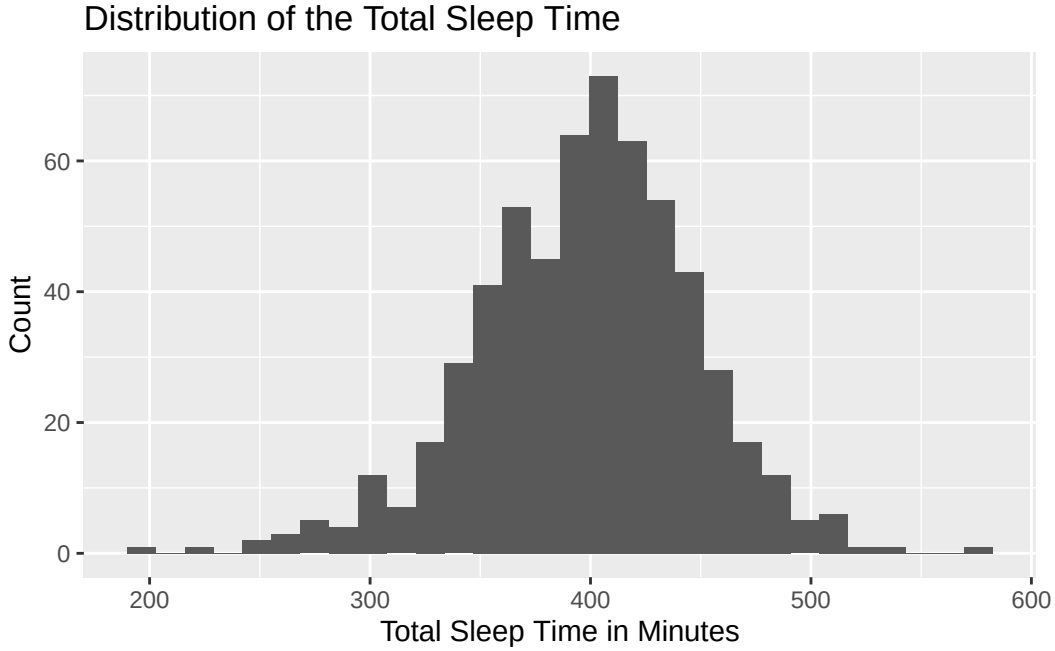


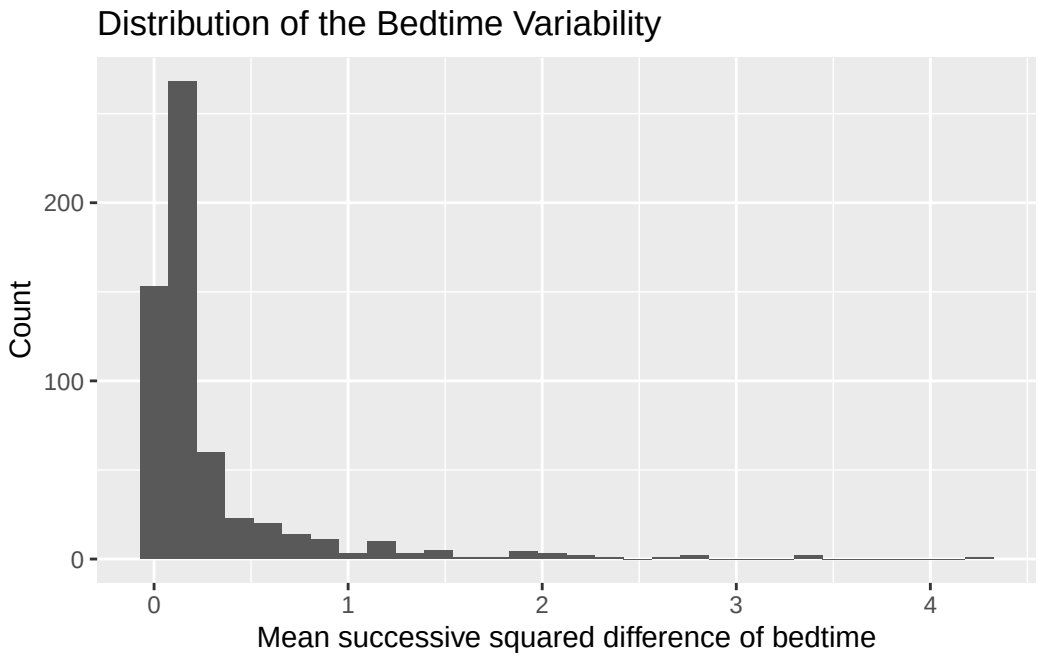


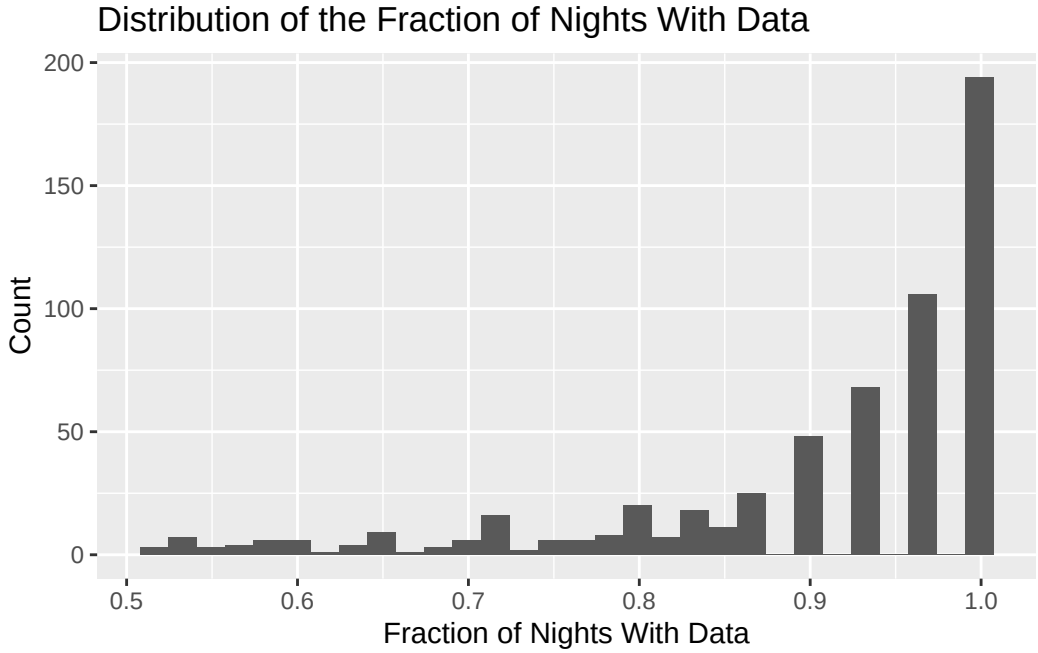

```
# A tibble: 3 x 7
  university mean_tgpa median_tgpa sd_tgpa min_tgpa max_tgpa count
  <chr>          <dbl>       <dbl>   <dbl>    <dbl>    <dbl> <int>
```

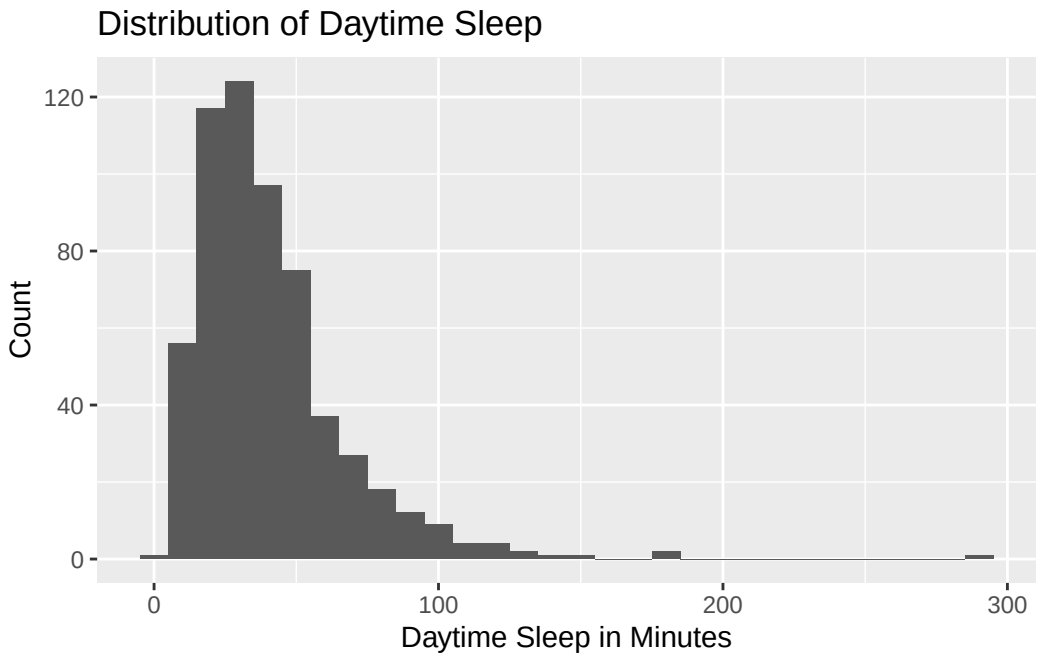
```
1 cath_priv       3.66        3.71   0.267     2.72        4   142
2 public          3.40        3.5    0.518     0.35        4   249
3 stem_priv       3.36        3.49   0.535     1.5         4   197
```

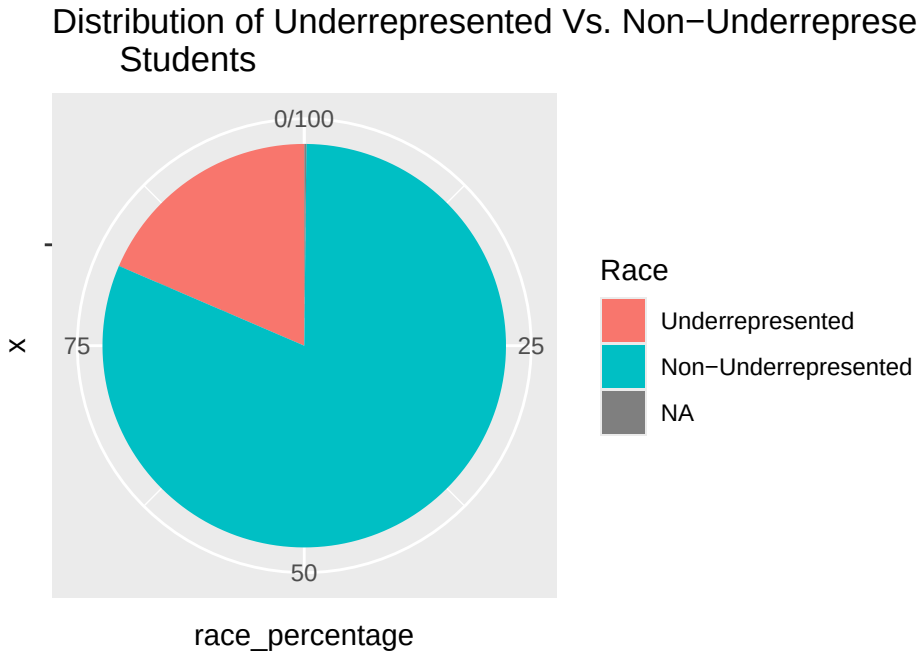
```
# A tibble: 3 x 7
  university mean_cgpa median_cgpa sd_cgpa min_cgpa max_cgpa count
  <chr>          <dbl>       <dbl>   <dbl>    <dbl>    <dbl> <int>
1 cath_priv       3.64        3.71   0.261     2.80        4   142
2 public          3.43        3.50   0.400     1.59        4   249
3 stem_priv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

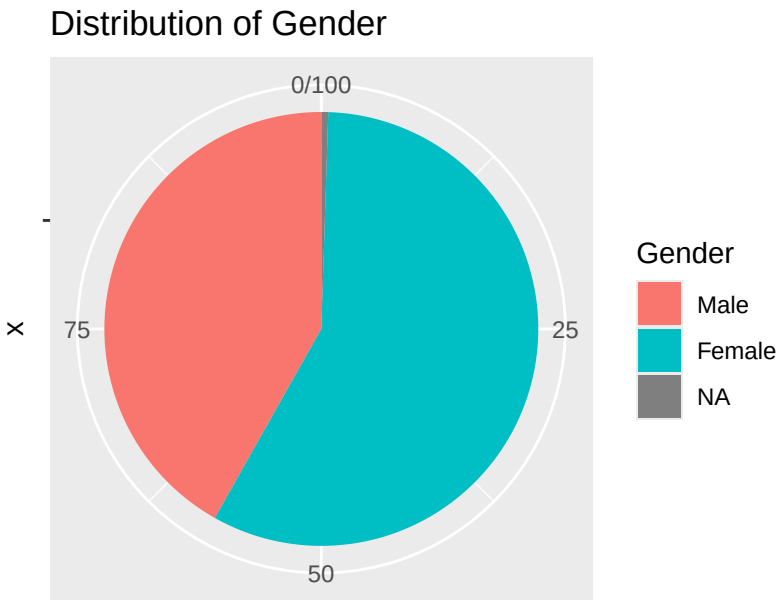


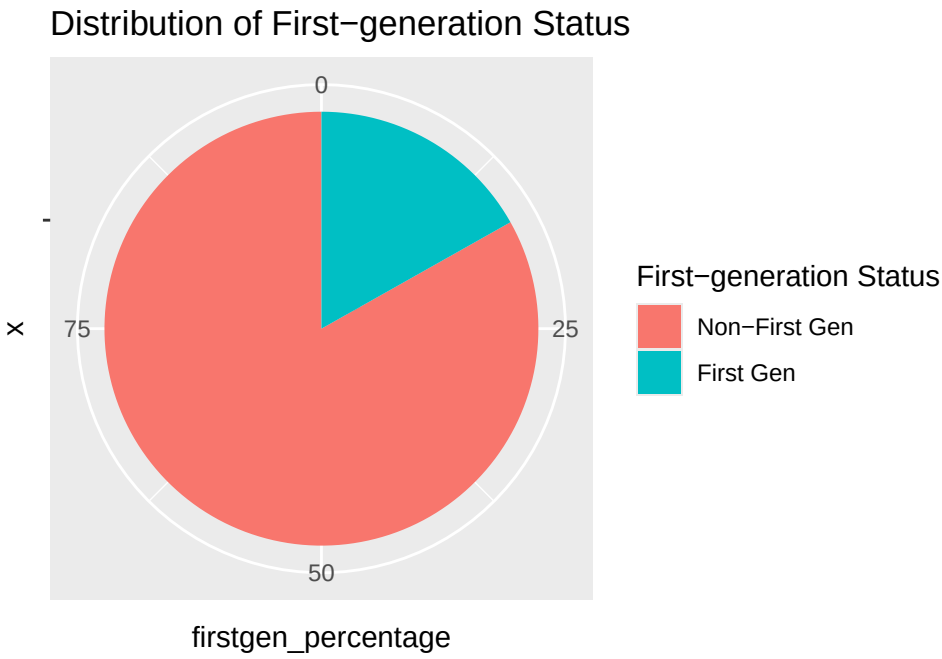

Distribution of the Bedtime Variability
Count
200
100
0
0
1
2
3
4
Mean successive squared difference of bedtime


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
75
25
50
x
race_percentage
Race
Underrepresented
Non−Underrepresented
NA

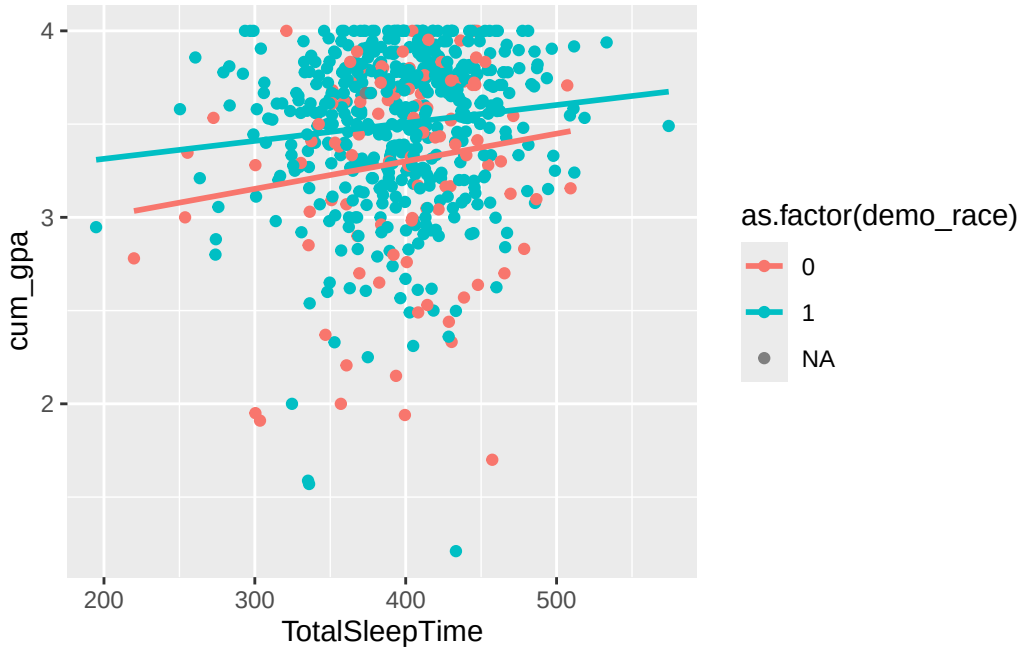


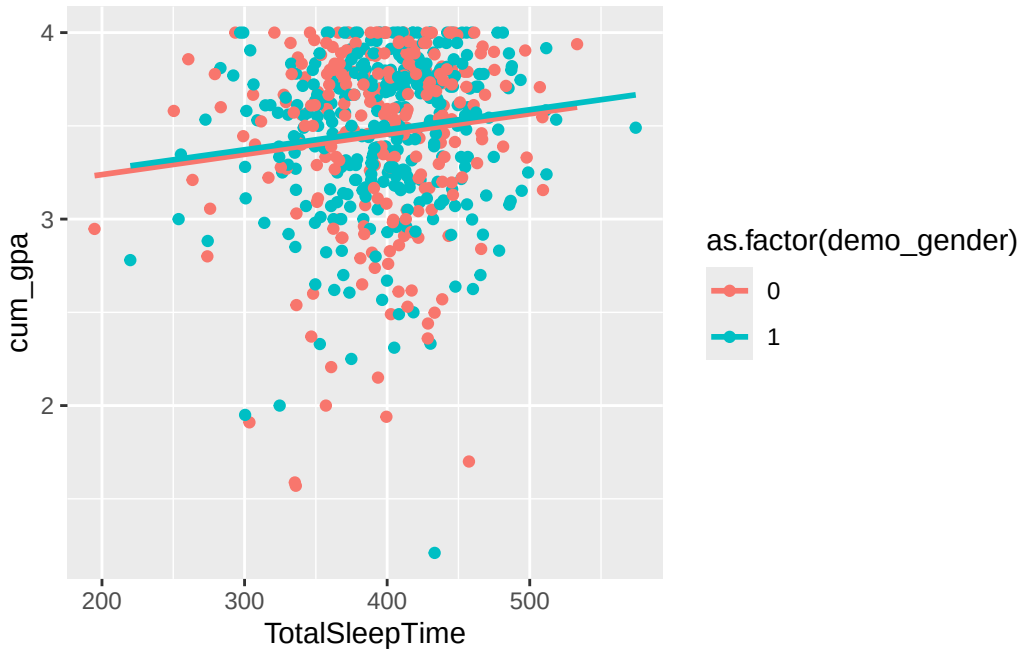


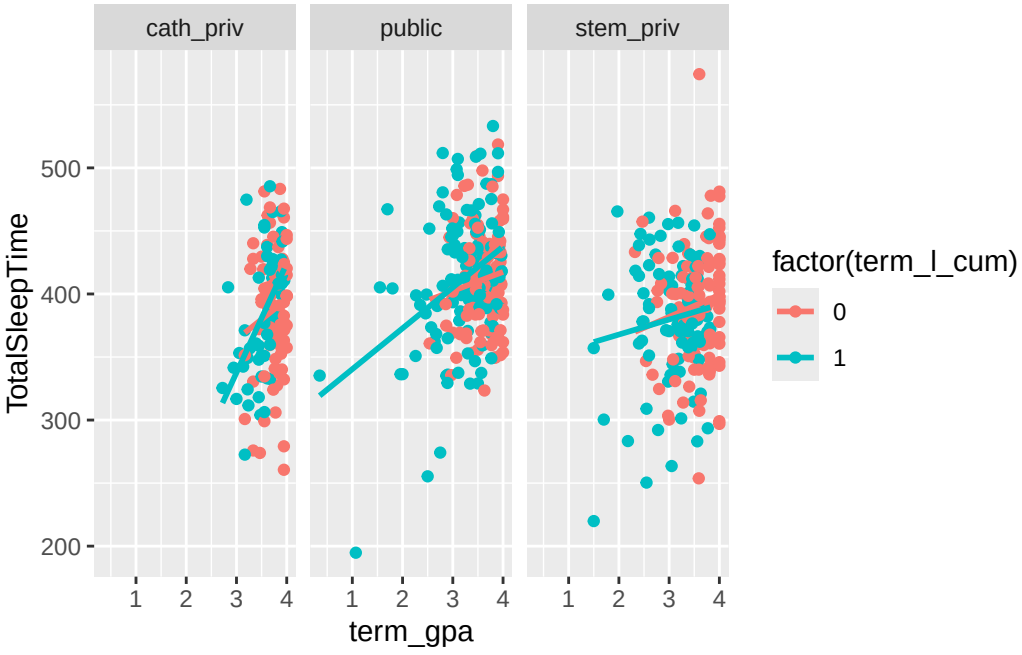

```
# A tibble: 3 x 4
  university total_count na_count non_na_count
  <chr>            <int>    <int>        <int>
```

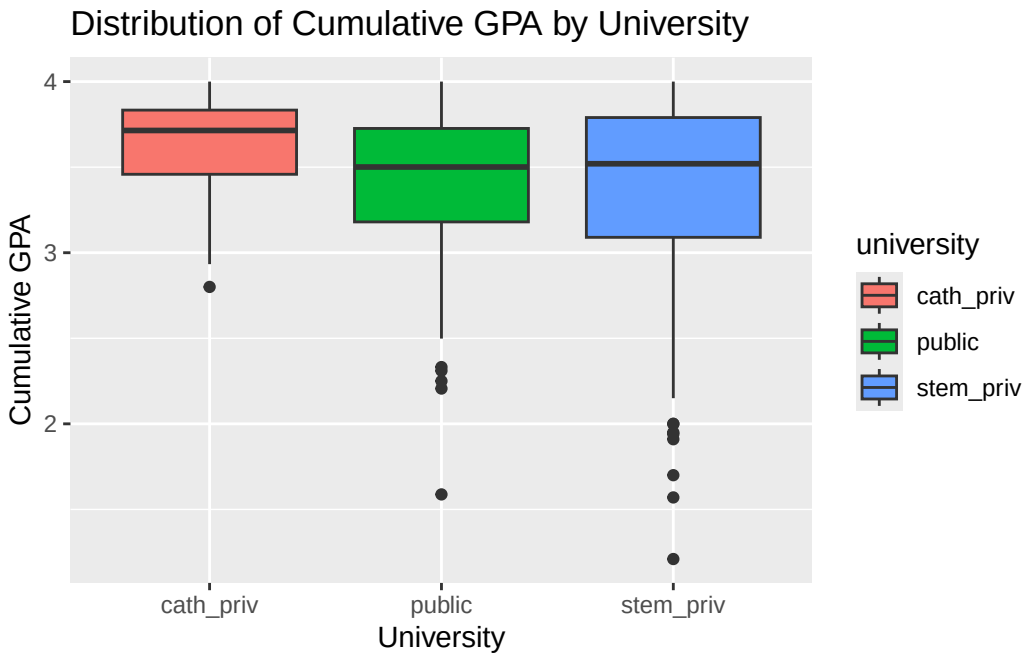
```
1 cath_priv          142       142             0
2 public             249         0           249
3 stem_priv          197         0           197
```

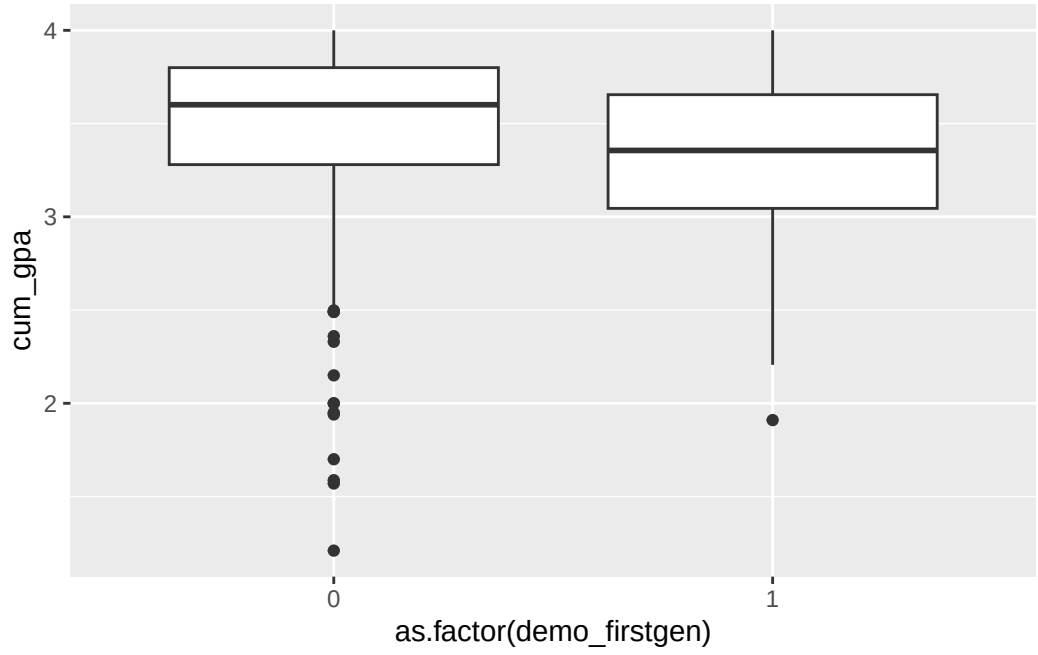
Bivariate EDA of The Response & Key Predictor Variables

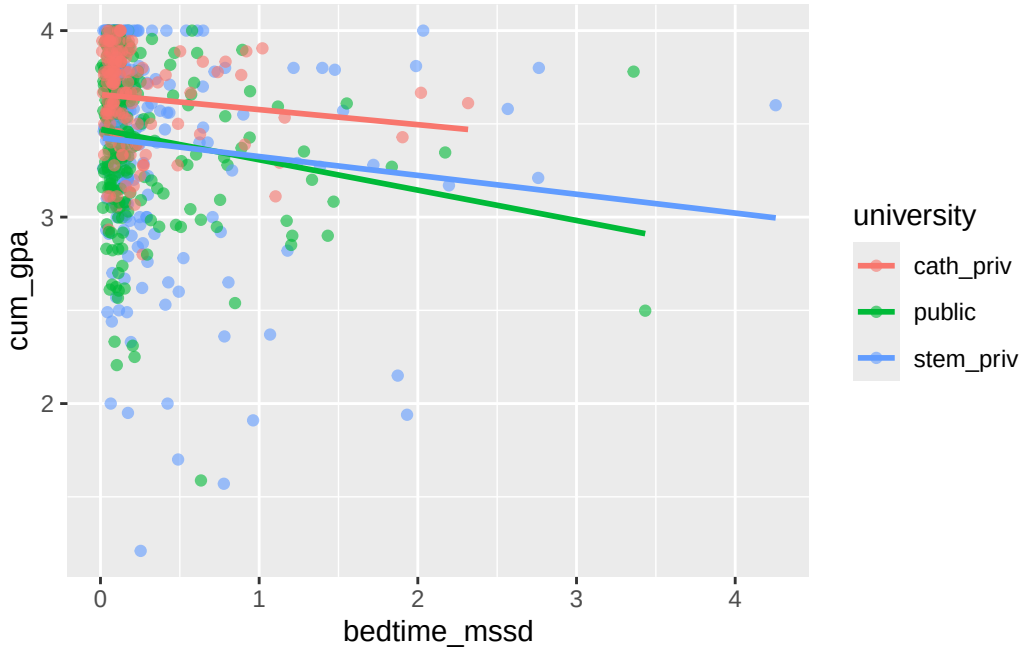


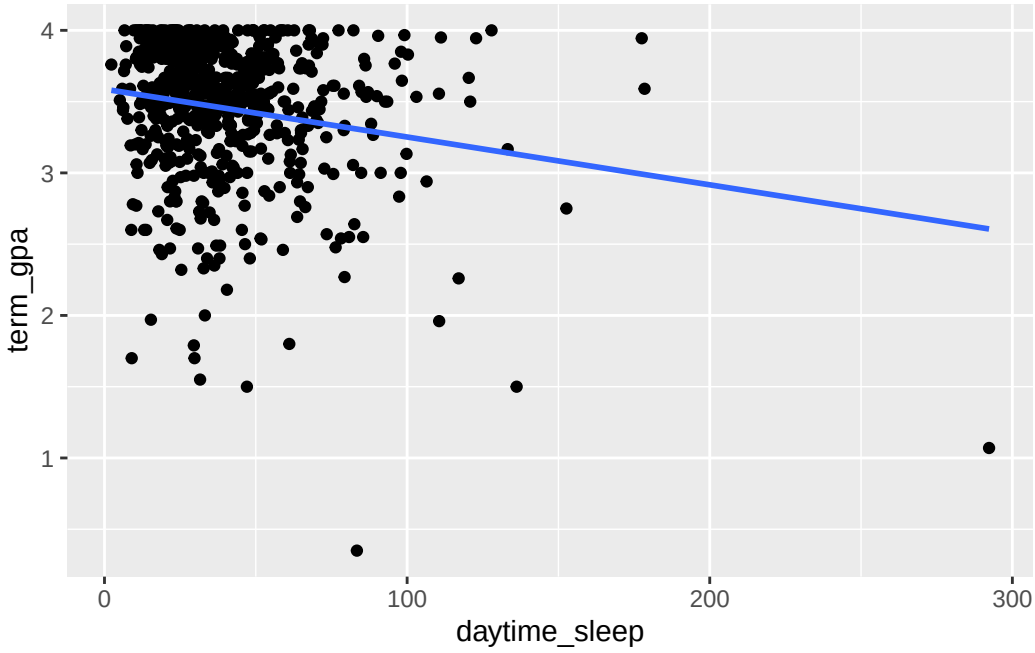

cum_gpa
4
3
2
200
300
400
500
TotalSleepTime
as.factor(demo_gender)
0
1


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


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


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





notes: stem_priv cluster is entirely lower than public cluster

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 differentiaed 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.