Project Proposal

WALE - Liane Ma, Amy Xu, Will Skelly, Eshan Mehere

```
library(tidyverse)
library(tidymodels)
library(dplyr)
library(knitr)

university_dataset <- read_csv("data/cmu-sleep.csv")
```

Introduction

Academic performance varies widely from student-to-student, likely attributed to a variety of factors including but not limited to students' amount of sleep, classes taken, university type, and background.

Our research question is: How do differences in sleep, race, gender, university type, and first-generation status affect college first-years' cumulative GPA?

It is important to determine what factors or combinations of factors can impact students' GPA, especially for first-years as they transition from high school to university. 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. It is well-known that sleep impacts students' academic achievement, 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. 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.

Data description

https://cmustatistics.github.io/data-repository/psychology/cmu-sleep.html

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.

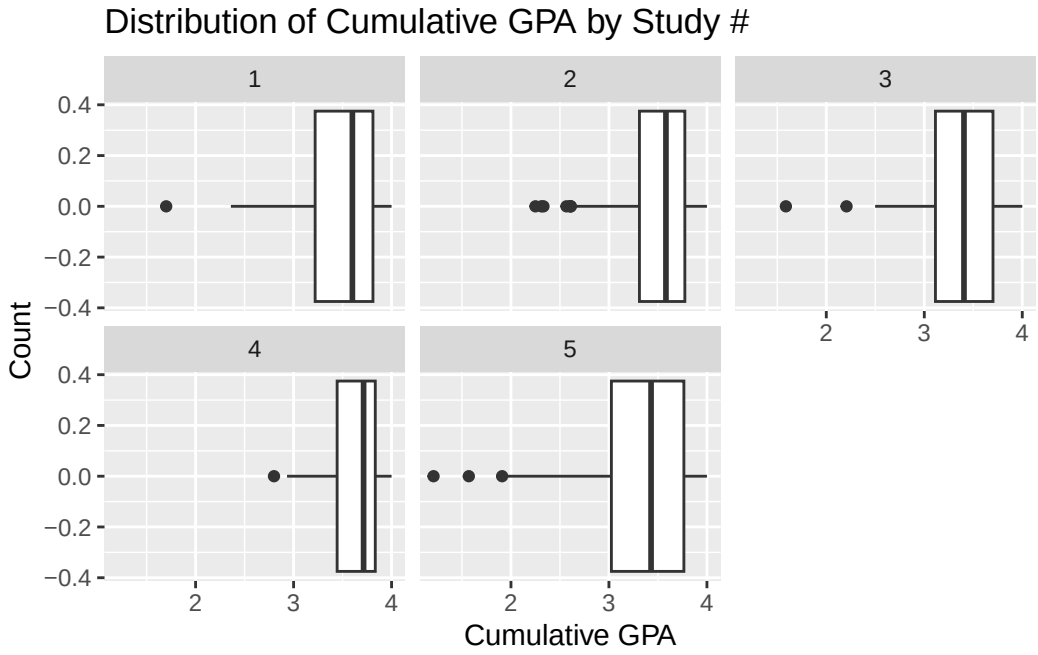
There are 634 observations, representing the 634 participants in this study. Race is a binary variable separated into underrepresented students and non-underrepresented students with 0 being underrepresented and 1 being non-underrepresented. Students are considered underrepresented if either parent is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

Exploratory data analysis

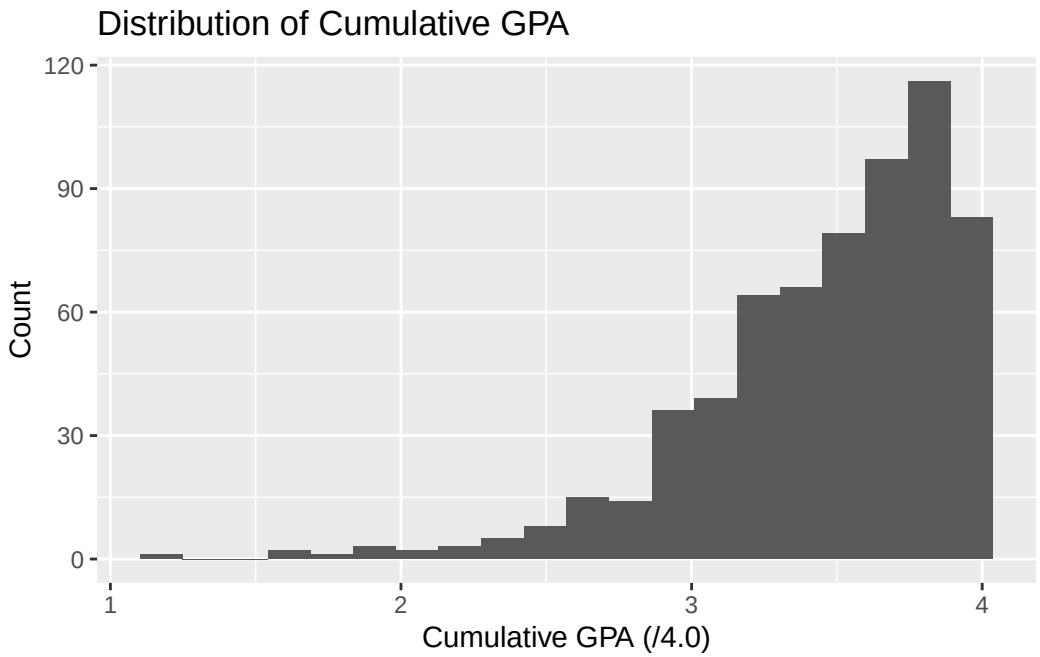
We need to clean the data by transforming the variable `study` to a variable for university type, where study 1 and 5 are data collected from a STEM-focused private university (CMU), study 2 and 3 are from a large public university (UW), and study 4 is from a small private Catholic university (ND). Additionally, we need to consider removing NAs or finding a way to replace them, or removing variables that are not needed that contain many NAs (i.e. `Zterm_units_ZofZ`.

Visualizations of the response variable:

```
university_dataset %>%
  ggplot(aes(
    x = cum_gpa
  )) +
  geom_boxplot() +
  facet_wrap(~study) +
  labs(
    title = "Distribution of Cumulative GPA by Study #",
    x = "Cumulative GPA",
    y = "Count"
  )
```




```
university_dataset %>%
  ggplot(aes(
    x = cum_gpa
  )) +
  geom_histogram(bins = 20) +
  labs(
    title = "Distribution of Cumulative GPA",
    x = "Cumulative GPA (/4.0)",
    y = "Count"
  )
```



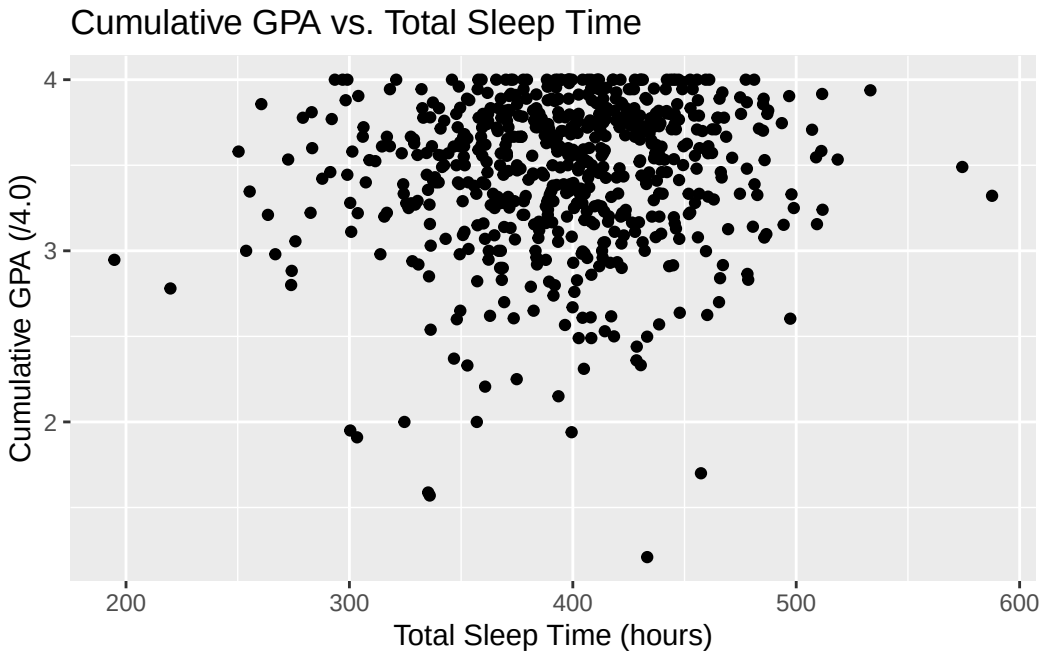

```
tst_summary <- university_dataset %>%
  summarize(
    mean = mean(TotalSleepTime, na.rm = TRUE),
    sd = sd(TotalSleepTime, na.rm = TRUE),
    min = min(TotalSleepTime, na.rm = TRUE),
    max = max(TotalSleepTime, na.rm = TRUE),
    median = median(TotalSleepTime, na.rm = TRUE)
  )
tst_summary %>%
  kable(caption = "Summary Statistics for Total Sleep Time (hours)")
```

Table 1: Summary Statistics for Total Sleep Time (hours)

mean	sd	min	max	median
397.3239	50.85673	194.7826	587.6667	400.3958

```
university_dataset %>%
  ggplot(aes(
    x = TotalSleepTime,
    y = cum_gpa
  )) +
```

```
  geom_point() +
  labs(
    title = "Cumulative GPA vs. Total Sleep Time",
    x = "Total Sleep Time (hours)",
    y = "Cumulative GPA (/4.0)"
  )
```



Some additional visualization options to explore this data in the future could include:

- Facet-wrapped/color-coded scatterplot by race for average time in bed vs. cumulative GPA
- Facet-wrapped/color-coded scatterplot by gender for average time in bed vs cumulative GPA
- Side by side boxplots (separated by university type) distribution of cumulative GPA
- Cumulative GPA vs. Term Units scatterplot
- Summary statistics of cumulative GPA
- Summary statistics/boxplot of totalsleeptime
- Scatterplot of daytime sleep vs. cumulative GPA

Analysis approach

Potential predictor variables of interest:

- bedtime_mssd
- TotalSleepTime
- daytime_sleep
- study
- term_units
- demo_gender
- demo_firstgen
- demo_race

We can use a multiple linear regression technique for this dataset that predicts GPA based on the above variables. We have both categorical and quantitative variables as predictors.

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]