

Student Sleep Patterns

Stat 184 Sec 2

Anna Kasehagen, Gracy Franco Prasanna, Melissa Kim, Sarah Khan

2024-12-13

1 Introduction

Understanding student sleep patterns and their relationship to physical activity is crucial for improving student well-being and academic performance. Sleep is an essential component of student health, yet it is often disrupted by the demands of college life. In this report, we explore various factors that may affect student sleep habits, and focus on potential differences across university years, physical activity levels, and the interaction between these two variables.

The data, from Kaggle,¹ is synthetically generated, meaning it was artificially created and does not represent real individuals. Although this synthetic data does not reflect real-world variance, it was intentionally designed to follow realistic distributions and relationships. This allows us to extract meaningful insights while still maintaining the benefits of controlled, consistent data patterns. It is worthy to note that real data may show more variability due to factors like randomness, self-reporting biases, and external influences on sleep and activity. While the analysis and conclusions based on this synthetic data provide valuable insights, caution should be taken when assuming that these findings directly apply to real-world data.

The data focuses on several key research questions. The cases in this data set represent hypothetical university students, with synthetic data on their sleep duration and physical activity levels. The main research questions explored are: 1.) How does sleep duration vary across the 4 university years? Similarly, how does physical activity vary across these years? 2.) How do physical activity levels differ between underclassmen (1st and 2nd years) and upperclassmen (3rd and 4th years) given their sleep duration? 3.) Is there a significant interaction between physical activity and sleep duration? 4.) What is the 5-number summary of sleep duration for all 4 university years?

These questions will guide further analysis, aiming to reveal meaningful trends and relationships within the data set.

2 FAIR/CARE Principles

Our data set, titled “Student Sleep Patterns” and created by Arsalan Jamal, meets the FAIR and CARE principles in the following manner:

¹jamal, A. (2024, October 14). Student sleep patterns. Kaggle. <https://www.kaggle.com/datasets/arsalanjamal002/student-sleep-patterns>

2.1 FAIR:

Findability: This data set is findable through the website Kaggle, which is an online platform that allows users to access data sets that can be used for a variety of data science and machine learning purposes. In addition, the name of this data set is “student_sleep_patterns.csv”, which succinctly describes the contents of the data set.

Accessibility: Kaggle is a highly accessible open-source platform, as its interface is user-friendly and its data sets are primarily free. Therefore, it provides a low barrier to entry for those interested in using its data sets. As such, this data set follows similar standards of accessibility. In our project, the data can be accessed through GitHub, which is also an open source platform, and, seeing as the repository is public, highly accessible.

Interoperability: The data is saved in a .csv format, which is widely-supported. In addition, common units are used, such as decimal points rounded to the tenths places.

Reusable: As this data is found through an open-source platform, which contains an extensive description of the data as well as the measurements of the data. In our project, we have created a QUARTO file with an introduction describing the data and our intentions with the data, as well as what we analysed.

2.2 CARE:

This data set is synthetic data, and as such, is neither perfectly reflective of student sleep patterns nor a comprehensive analysis of all the factors that contribute to student sleep patterns. This data, therefore, does not reflect the sleep patterns of any racial or ethnic group in particular, and does not directly draw from the data of Indigenous communities. This data also does not assess confounding factors that could, in a real-world setting, influence Indigenous students’ sleep schedules, such as systemic barriers or cultural practices.

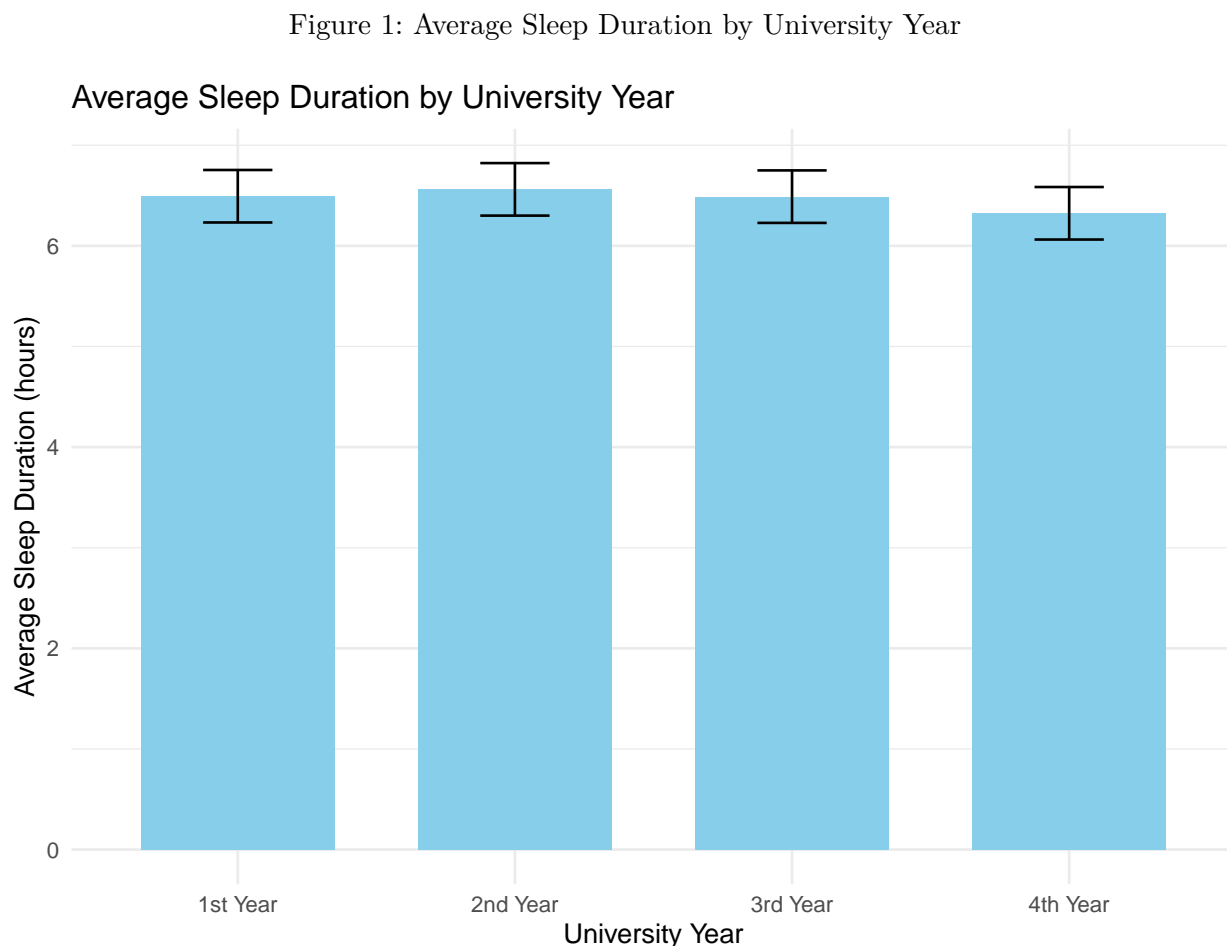
3 Data Exploration and Research Questions

3.1 Research Question 1

To begin our analysis, we examine the first research question: How does sleep duration vary across the four university years, and how does physical activity compare across these same groups? Understanding these trends can provide insight into how student lifestyles evolve throughout their time in university, reflecting changes in academic workload, priorities, and habits.

The bar graph of average sleep duration by university year (Figure 1) reveals a gradual decline in sleep as students advance through their studies. First-year students report the highest average sleep duration, approximately 6.5 hours, while fourth-year students report slightly less, averaging just over 6 hours. The error bars for each year are relatively narrow, indicating consistent sleep patterns within each group. This decrease may be attributed to increasing academic demands, extracurricular commitments, or stressors commonly associated with higher years of study. Despite the modest decline, the results suggest that sleep duration remains relatively stable across all years, with no drastic reductions observed.

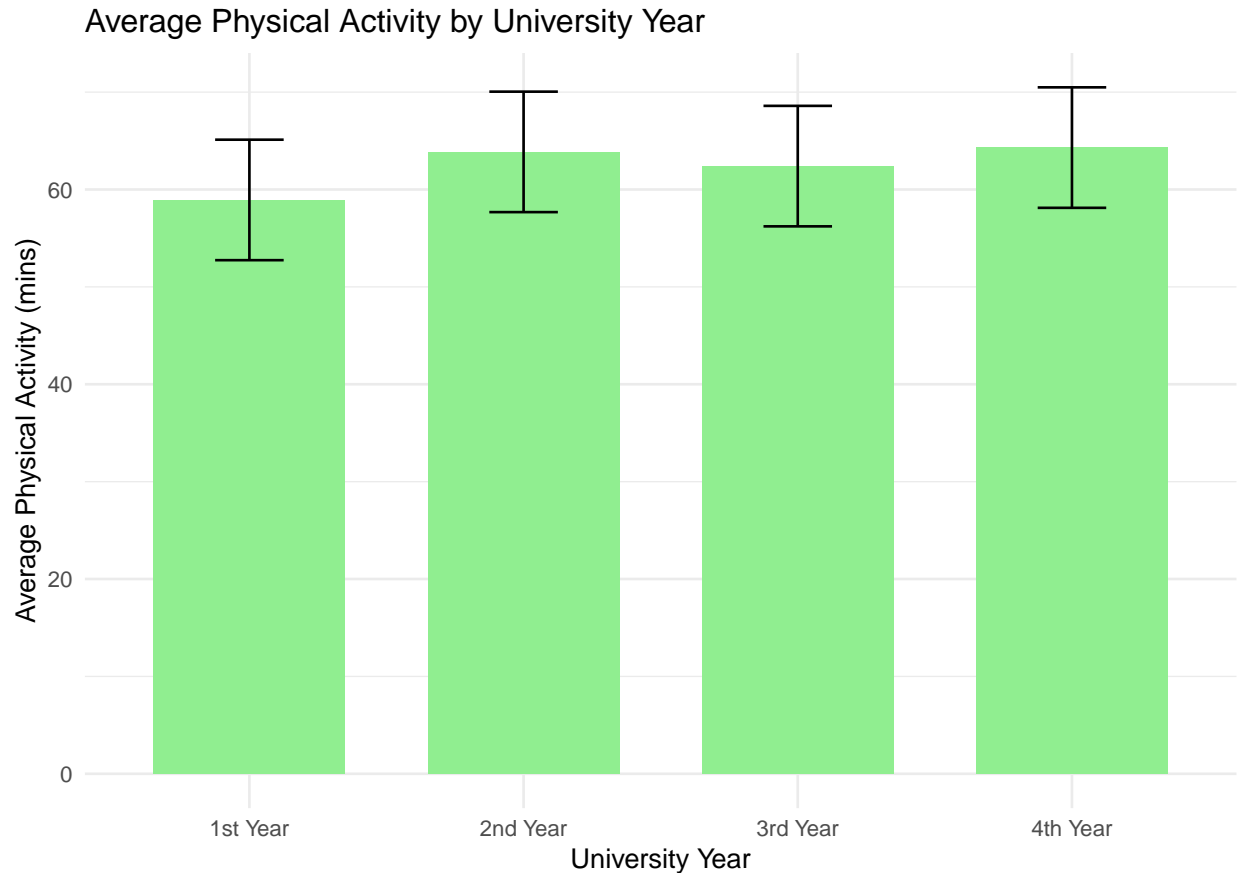
Figure 1: Bar Graph of Average Sleep Duration by University Year



In contrast, the bar graph for average physical activity (Figure 2) shows an opposite trend, with activity levels increasing as students progress through university. First-year students demonstrate the lowest levels of physical activity, averaging a little under 60 minutes, whereas fourth-year students report the highest levels, nearing 65 minutes. The error bars for physical activity are noticeably wider, particularly in the first two years, suggesting greater variability in exercise habits among younger students. This variability may stem from differences in schedules, personal habits, or adaptation to university life. By the later years, physical activity appears to stabilize, with upperclassmen engaging in more consistent and higher levels of activity. This trend may reflect a growing awareness of health and well-being or improved time management skills among older students.

Figure 2: Bar Graph of Average Physical Activity Level by University Year

Figure 2: Average Physical Activity by University Year



Together, these graphs highlight a subtle but interesting shift in student behavior over their university years. As sleep duration decreases slightly, physical activity tends to increase, indicating a potential shift in priorities or coping mechanisms. While the synthetic nature of the data ensures controlled patterns, these results align with the common narrative that upperclassmen, despite facing greater academic pressures, may adopt healthier habits such as increased physical activity to manage stress and maintain balance.

3.2 Research Question 2

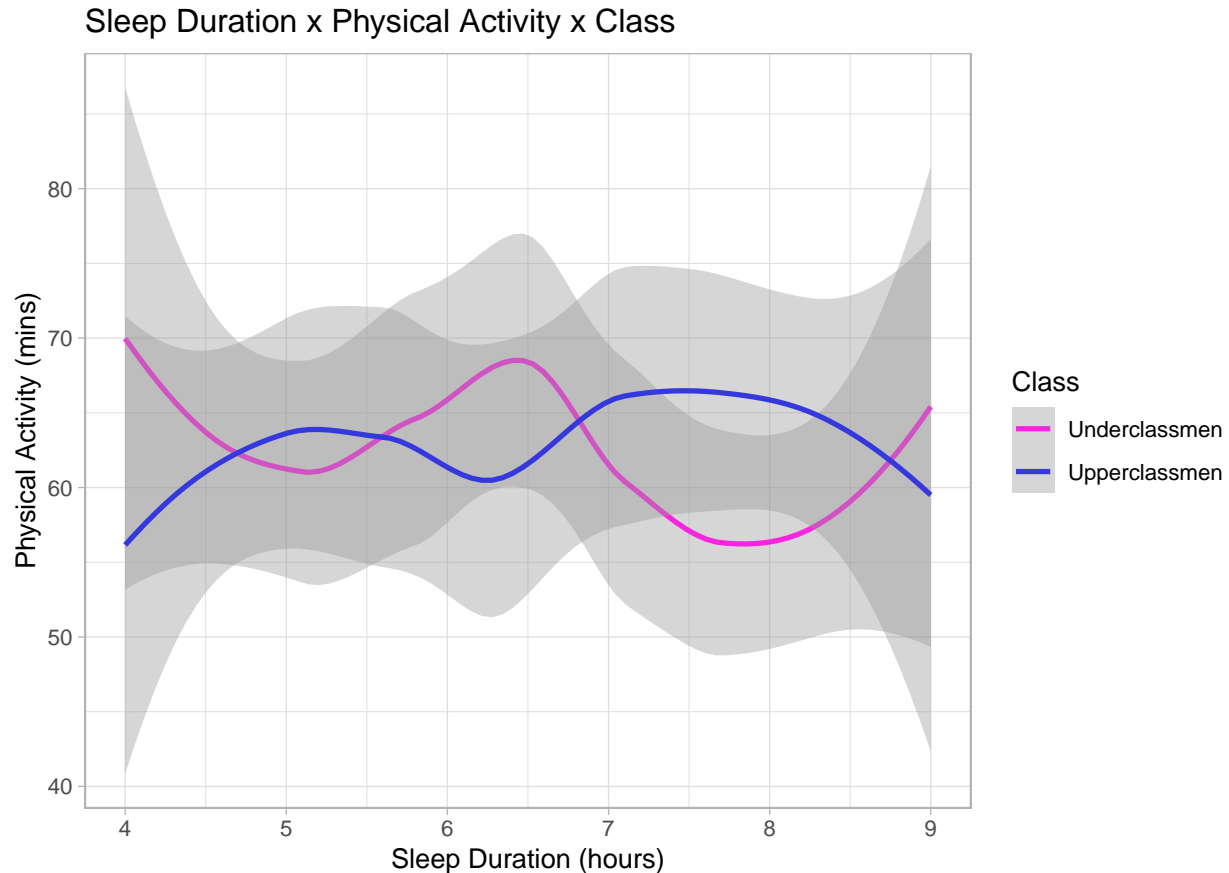
After analyzing how sleep duration and physical activity vary by university years, we wanted to see how given two of the attributes for a particular student, how the third would be affected. More specifically, the question we are trying to answer is:

- Given the class group of a student (underclassmen or upperclassmen) and their sleep duration, how do their reported levels of physical activity vary?

To best visualize this relationship, we thought that a line graph would work best, and that way we would be able to see the trend between the two classes with duration of sleep spanning from 4 to 9 hours. The graph below demonstrates this relationship. Some things to keep in mind are that the x axis, sleep duration, is measured in hours, while the y axis, physical activity levels, are measured

in minutes, and underclassmen are represented by the pink line and upperclassmen are represented by the blue line. The gray bubbles around the lines represent the confidence intervals, but for our analysis we just focused on the average lines themselves.

Figure 3: How Physical Activity Levels vary with Sleep Duration and Class Group



From this graph, there were a few things we noticed and found interesting:

- The first is how the trends for upperclassmen and underclassmen seem to be inverted. That is, how at a given hour for sleep duration, the physical activity of one group goes up while the other goes down, or vice versa.
- Second, if we look at where the lines cross the 8 hour mark (we chose 8 because it is right in the middle of the recommended amount of sleep a person should get every night (7-9)²) we can see a large gap between the recorded minutes of physical activity an underclassman might get (just above 55 mins) versus an upperclassman (just above 65 mins).
- The third interesting takeaway from looking at this graph is how we could visually see at what point of sleep duration do upperclassmen and underclassmen hit their maximum physical

²Assess your sleep needs. Sleep Medicine. (n.d.). <https://sleep.hms.harvard.edu/education-training/public-education/sleep-and-health-education-program/sleep-health-education-92#:~:text=Although%20there%20is%20some%20genetic,sleep%20as%20long%20as%20possible>.

activity levels. We see that for underclassmen it is at 4 hours whereas for upperclassmen it is at about 7.5/8 hours.

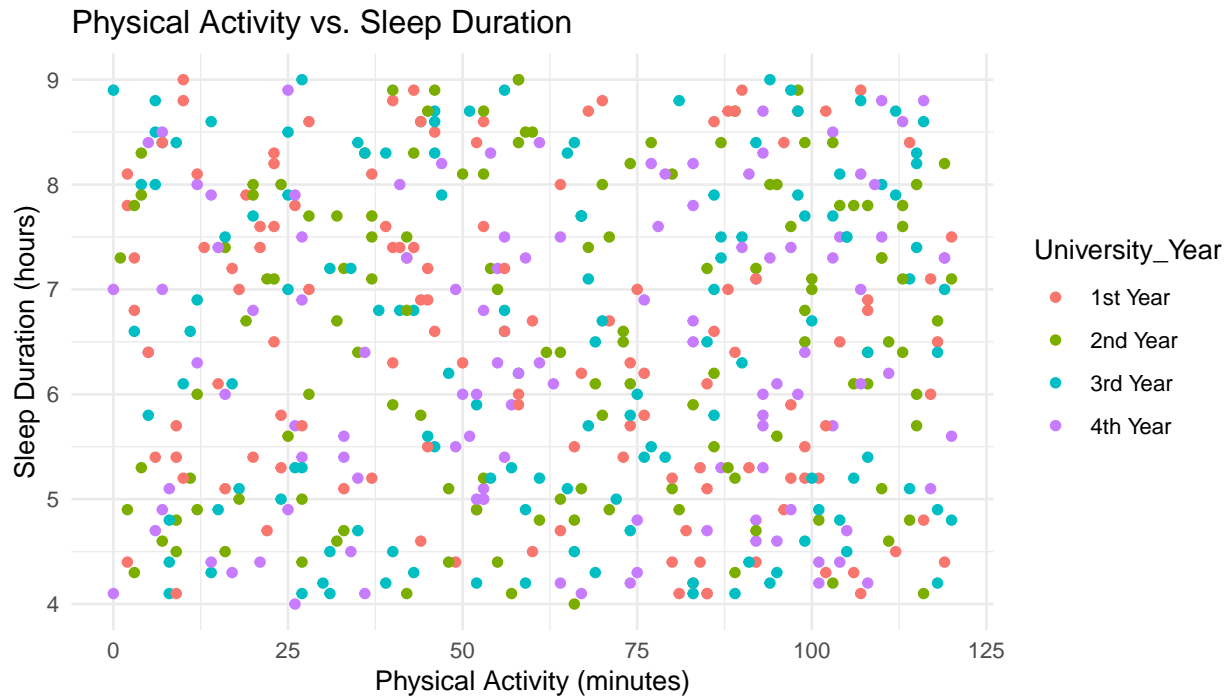
After considering these takeaways, we can make inferences about what might have influenced these sleep patterns and activity levels between classes. The jump between physical activity levels at certain hours and the difference between where the groups reached their maximum levels of sleep duration could be attributed to a difference in time management skills. As students reach their 3rd and 4th years, they typically have figured out a schedule that allows them to balance getting good sleep while maintaining an active lifestyle. This trend could also be attributed to how upperclassmen might have learned the value of maintaining their physical health, while underclassmen might be spreading themselves thin since they have just entered a new environment.

3.3 Research Question 3

We wanted to create a wholistic analysis regarding the correlation of physical activity and students' sleep duration, regardless of student year. To find this, we proposed the following research question:
- Is there a significant interaction between physical activity and sleep duration?

To best analyze this, we created a dot plot to show the distribution of students' physical activity in minutes (the independent variable found on the 'x' axis) and students' sleep duration in hours (the dependent variable found on the 'y' axis). Each dot will also be one of four colors, each color representing a university year: red for first years, blue for second years, green for third years, and purple for fourth years. This information can also be found in the key of the data visualization below.

Figure 4: The Correlation between Physical Activity and Sleep Duration Across All Student Years



From this visualization, we were able to gain a few key insights:

- There seems to be no significant correlation between students' physical activity levels and sleep duration. This could be due to a number of confounding factors -such as class schedule, workload, major, personal events, lifestyle choices- that are not accounted for within the data set.
- In addition, this could also be due to the synthetic nature of the data set we are analyzing, which, even if it draws from existing measured attributes of sleep duration and physical activity, might not be a perfect representation of these variables, and may skew our findings and their applicability to real-world scenarios.

3.4 Research Question 4

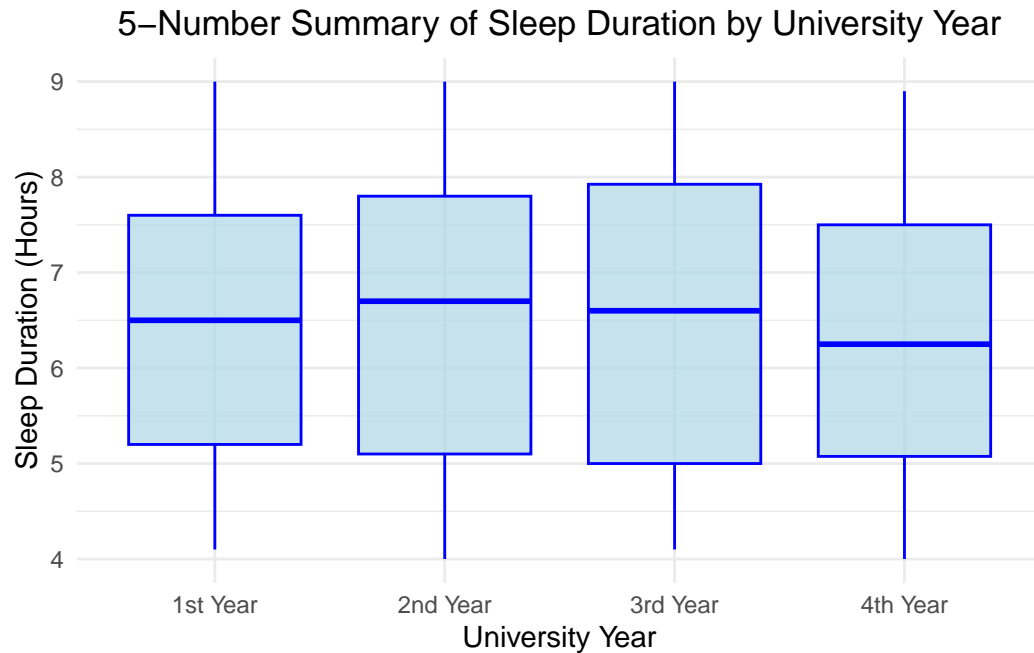
In this final visualization, we are concerned with seeing how two attributes (University Year and Sleep Duration) are related and potentially seeking out any correlations between the two. In the previous visualizations, the focus has primarily been looking at general averages across university years or a direct comparison between a group of classes, but here we want to see a statistical summary to include other measures of central tendency and ranges between the university classes.

- Is there a trend in sleep duration as students progress through university years (e.g., does it decrease/vary in later years)?
- What factors might contribute to changes in sleep duration as students advance through their academic journey?

To best visualize this relationship, we thought that a side by side comparison of a box plot with the following five measures: Minimum, Q1, Median, Q3, and Maximum, would be best. The x axis holds the university year while the y axis holds sleep duration. The measures are demarcated by the dark blue outlines, while the light blue highlights where 50% of the data closest to the median lies. The box plot is important to understand where the majority of people lie within the class, and allows for a more wholistic comparison between the multiple university classes.

Table 1: Five Number Summary

University_Year	Min	Q1	Median	Q3	Max
1st Year	4.1	5.200	6.50	7.600	9.0
2nd Year	4.0	5.100	6.70	7.800	9.0
3rd Year	4.1	5.000	6.60	7.925	9.0
4th Year	4.0	5.075	6.25	7.500	8.9



A few key insights to take from here:

- The median sleep duration does not vary much across the years, central tendency is fairly consistent for all university students.
- For second and third year students, the range is more narrow, so they tend to have more consistency in their sleep patterns compared to first and fourth years.
- There are no significant outliers in the data set, which means there is consistency among most students in their respective cohort.
- The second and third year students likely have stability in their workloads and schedules, meaning the ranges are narrower. For first or third year students it can be different due to initial adjustments or final years containing a wider variety of workloads due to the nearing of graduation or finding a job.

4 Conclusion

In this project, we took synthetic data into consideration to see how a student's sleep patterns change throughout years of university. Our results showed nuanced and intercorrelated insights to both student life, and how it may change and ultimately affect a student's quality of sleep.

Sleep Patterns

As a student transitions to the next year, we see general decreases in average sleep duration. This gradual decline is likely attributed to greater academic obligations or potentially demanding lifestyle changes and prioritization.

Physical Activity Trends

Physical activity was, interestingly enough, higher in upperclassmen than underclassmen. This may indicate a greater sense of time management and school-life balance that comes with experience in college.

Beyond the Numbers

While there is not evidently a strong correlation between sleep duration and physical activity, our visualizations and research suggests that individual factors like lifestyle choices, priorities of academia over other responsibilities, and intrapersonal skills like time management likely have a more significant impact on the way students go about their lives. It sheds light on how undergraduate students' health behaviors subtly differ throughout their journey in school, emphasizing that the interplay between academic life and personal life is a unique and important balance to strike for overall well-being.

5 Code Appendix

```
#loading necessary packages
library(tidyr)
library(dplyr)
library(rvest)
library(google sheets4)
library(ggplot2)
library(esquisse)
library(tidyverse)

#reading in the data set
sleepData <- read_sheet("https://docs.google.com/spreadsheets/d/1BszLI2k3ti0AzKrY6msUY5lqb1DG-")

#selects just the columns applicable for our research questions
CleanedSleepData <- sleepData %>%
  select(4, 5, 9) %>%
  na.omit() #checks for and ignores any values that have NA

# Group data by University_Year and calculate the average Sleep_Duration and Physical_Activity
averages <- CleanedSleepData %>%
  group_by(University_Year) %>%
  summarise(
    avg_sleep_duration = mean(Sleep_Duration, na.rm = TRUE),
    avg_physical_activity = mean(Physical_Activity, na.rm = TRUE)
  )

# Calculate Confidence Intervals for Sleep Duration and Physical Activity for each University
conf_intervals <- CleanedSleepData %>%
  group_by(University_Year) %>%
  summarise(
    sleep_duration_ci = list(t.test(Sleep_Duration)$conf.int),
```

```

    physical_activity_ci = list(t.test(Physical_Activity)$conf.int)
  )

# Calculate the percentage change in average sleep duration and physical activity between years
averages <- averages %>%
  arrange(University_Year) %>%
  mutate(
    sleep_duration_pct_change = c(NA, diff(avg_sleep_duration) / head(avg_sleep_duration, -1)) * 100,
    physical_activity_pct_change = c(NA, diff(avg_physical_activity) / head(avg_physical_activity, -1)) * 100
  )

# Join the averages with the confidence intervals to add the confidence interval values
averages_with_ci <- averages %>%
  left_join(conf_intervals, by = "University_Year")

# Bar plot for Sleep Duration
ggplot(averages_with_ci, aes(x = University_Year, y = avg_sleep_duration)) +
  geom_bar(stat = "identity", fill = "skyblue", width = 0.7) +
  geom_errorbar(aes(
    ymin = avg_sleep_duration - (sleep_duration_ci[[1]][2] - sleep_duration_ci[[1]][1]) / 2,
    ymax = avg_sleep_duration + (sleep_duration_ci[[1]][2] - sleep_duration_ci[[1]][1]) / 2
  ), width = 0.25) +
  labs(title = "Average Sleep Duration by University Year", y = "Average Sleep Duration (hours)") +
  theme_minimal()

#loading necessary packages
library(tidyr)
library(dplyr)
library(rvest)
library(google sheets4)
library(ggplot2)
library(esquisse)
library(tidyverse)

#reading in the data set
sleepData <- read_sheet("https://docs.google.com/spreadsheets/d/1BszLI2k3ti0AzKrY6msUY5lqblDG-c")

#selects just the columns applicable for our research questions
CleanedSleepData <- sleepData %>%
  select(4, 5, 9) %>%
  na.omit() #checks for and ignores any values that have NA

# Group data by University_Year and calculate the average Sleep_Duration and Physical_Activity
averages <- CleanedSleepData %>%
  group_by(University_Year) %>%
  summarise(
    avg_sleep_duration = mean(Sleep_Duration, na.rm = TRUE),
    avg_physical_activity = mean(Physical_Activity, na.rm = TRUE)
  )

```

```

)

# Calculate Confidence Intervals for Sleep Duration and Physical Activity for each University
conf_intervals <- CleanedSleepData %>%
  group_by(University_Year) %>%
  summarise(
    sleep_duration_ci = list(t.test(Sleep_Duration)$conf.int),
    physical_activity_ci = list(t.test(Physical_Activity)$conf.int)
  )

# Calculate the percentage change in average sleep duration and physical activity between years
averages <- averages %>%
  arrange(University_Year) %>%
  mutate(
    sleep_duration_pct_change = c(NA, diff(avg_sleep_duration) / head(avg_sleep_duration, -1)) * 100,
    physical_activity_pct_change = c(NA, diff(avg_physical_activity) / head(avg_physical_activity, -1)) * 100
  )

# Join the averages with the confidence intervals to add the confidence interval values
averages_with_ci <- averages %>%
  left_join(conf_intervals, by = "University_Year")

# Bar plot for Physical Activity
ggplot(averages_with_ci, aes(x = University_Year, y = avg_physical_activity)) +
  geom_bar(stat = "identity", fill = "lightgreen", width = 0.7) +
  geom_errorbar(aes(
    ymin = avg_physical_activity - (physical_activity_ci[[1]][2] - physical_activity_ci[[1]][1]),
    ymax = avg_physical_activity + (physical_activity_ci[[1]][2] - physical_activity_ci[[1]][1]),
    width = 0.25) +
  labs(title = "Average Physical Activity by University Year", y = "Average Physical Activity") +
  theme_minimal()

#changes the names of the columns so that 1st and 2nd years are underclassmen and 3rd and 4th years are upperclassmen
Q2SleepData <- CleanedSleepData %>%
  mutate(
    University_Year = case_match(
      .x = University_Year,
      "1st Year" ~ "Underclassmen",
      "2nd Year" ~ "Underclassmen",
      "3rd Year" ~ "Upperclassmen",
      "4th Year" ~ "Upperclassmen",
      .default = "missing"
    )
  )

# Make Data Visualization ----
ggplot(
  data = Q2SleepData,
  mapping = aes(

```

```

    x = Sleep_Duration,
    y = Physical_Activity,
    colour = University_Year
  )
) +
geom_smooth(se = TRUE) +
scale_color_manual(
  values = c(Underclassmen = "#F725DF",
             Upperclassmen = "#3538DC")
) +
labs(
  x = "Sleep Duration (hours)",
  y = "Physical Activity (mins)",
  title = "Sleep Duration x Physical Activity x Class",
  color = "Class"
) +
theme_light()
# Make Data Visualization ----
ggplot(CleanedSleepData) +
  aes(
    x = Physical_Activity,
    y = Sleep_Duration,
    colour = University_Year
  ) +
  geom_point() +
  scale_color_hue(direction = 1) +
  theme_minimal() +
  labs(
    x = "Physical Activity (minutes)",
    y = "Sleep Duration (hours)",
    title = "Physical Activity vs. Sleep Duration")
library(dplyr)
library(kableExtra)
# Getting the stats for the five number summary (for each year)
SleepSummary <- CleanedSleepData %>%
  group_by(University_Year) %>%
  summarize(
    Min = min(Sleep_Duration, na.rm = TRUE),
    Q1 = quantile(Sleep_Duration, 0.25, na.rm = TRUE),
    Median = median(Sleep_Duration, na.rm = TRUE),
    Q3 = quantile(Sleep_Duration, 0.75, na.rm = TRUE),
    Max = max(Sleep_Duration, na.rm = TRUE)
  )

#uses kable package to make table easier to read
SleepSummary %>% kable(
  caption = "Five Number Summary", #adds a caption

```

```

booktabs = TRUE,
align = c("l", rep("c", 6))
) %>%
kableExtra::kable_styling(
bootstrap_options = c("striped", "condensed"), #stripes and condenses the table for easier read
font_size = 16 #assigns a readable font size
)

# Print the 5-number summary for personal accuracy
#print(SleepSummary)

# Creating the Boxplot visualization for Sleep Duration by University Year
ggplot(CleanedSleepData, aes(x = factor(University_Year), y = Sleep_Duration)) +
  geom_boxplot(fill = "lightblue", color = "blue", alpha = 0.7) +
  labs(
    title = "5-Number Summary of Sleep Duration by University Year",
    x = "University Year",
    y = "Sleep Duration (Hours)"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

```