Project

STAT 231: Calendar Query

Justin Papagelis

Last updated March 11, 2022

Introduction

The questions that I choose to focus on are as follows:

- What is the average amount of time I spend studying during the week? How much of this time is spent on each class?
- How much time do I spend studying in relation to sleep.

These questions are important to me because I always try to develop healthy study and sleep habits, so by collecting my data and analyzing it, I can make sure that I am staying healthy. This project will also be helpful to me so that I can change any unhealthy habits if necessary. It will also be interesting to see the breakdown of how I spend my time studying.

Data collection

I collected data using an Apple Calendar by inputting the chunks of time when I was studying or sleeping. My variables of interest were which activity I was doing and the duration of that activity. For which activity I was doing, I focused on sleeping and my four Amherst College courses (Mixed-Race America, Theoretical Statistics, Data Science, and Networks). For the duration of the activity, the intervals were recorded using hours. Since I used the calendar to track my activities, the exact date of each activity was also recorded. This was helpful, so I could use the day of the week as another categorical variable (with levels being Monday, Tuesday, etc.).

I will create a plot that shows how much time I spend studying compared to sleeping each day that I recorded my activity. The explanatory variable will be the date, and the response variables will be the total time I spend sleeping and studying each day. Some visual cues are the position of the point regarding the axes and the point's color to determine which activity. I will also create a bar plot to see the average amount of time I spend studying per day. Color will be used to distinguish between classes. I will create a table to show the average amount of time I spend doing each activity for one session (I define a session as a time interval in which I did one activity). This table will include summary statistics such as total time and mean.

```r
# Data import and preliminary wrangling
calendar_data <- "my_data.ics" %>%
  ## Use ical package to import into R
  ical_parse_df() %>%
  ## Convert to "tibble" data frame format
  as_tibble() %>%
  ## calendar event descriptions are in a variable called "summary"
  ## "activity" is a more relevant/informative variable name
  rename(activity = summary) %>%
  mutate(
    ## Specify time zone (defaults to UTC otherwise)
    start_datetime = with_tz(start, tzone = "America/New_York"),
    end_datetime = with_tz(end, tzone = "America/New_York"),
    ## Compute duration of each activity in hours
    duration = interval(start_datetime, end_datetime) / hours(1),
    ## Convert text to lower case and trim spaces to help clean up
    ## potential inconsistencies in formatting
    activity = str_to_lower(activity),
    ## separate date from time
    date = floor_date(end_datetime, unit = "day"),
    ## Examples of ways to parse dates, times
    year = year(date),
    month = month(date, label = FALSE),
    day = day(date),
    day_of_week = wday(date, label = TRUE),
    day_of_year = yday(date)) %>%
  ## remove spurious year (added to every Google calendar)
    filter(year != 1969)
```

```r
# wrangling on the large dataset
sleep <- calendar_data %>%
  # keep just the activities that correspond to sleep
  filter(activity == "sleep") %>%
  # calculate the total amount of time spent in each class for each day
  group_by(date) %>%
  mutate(total_sleep = sum(duration)) %>%
  # remove duplicates
  distinct(date, .keep_all = TRUE) %>%
  # rename
  rename(Sleep = total_sleep) %>%
  # just keep the variables date and Sleep
  select(date, Sleep)

study <- calendar_data %>%
  # keep just the activities that correspond to classes
  filter(activity != "sleep") %>%
  # calculate the total amount of time spent in each class for each day
  group_by(date) %>%
  mutate(total_study = sum(duration))  %>%
  # remove duplicates
  distinct(date, .keep_all = TRUE) %>%
  # rename
  rename(Study = total_study) %>%
```

```
  # just keep the variables date and Study
  select(date, Study)

# combine the two datasets
sleep_and_study <- sleep %>%
  left_join(study, by = "date") %>%
  # lengthen data
  pivot_longer(cols = 2:3, names_to = "activity", values_to = "time") %>%
  # gave the nas a value of 0
  replace_na(list(time = 0))
```

```
# wrangling on the large dataset
classes <- calendar_data %>%
  # only keep the class activities
  filter(activity != "sleep") %>%
  # calculate the total time spent on a class for a specific date
  group_by(date, activity) %>%
  mutate(sum = sum(duration)) %>%
  # remove duplicated dates
  distinct(date, .keep_all = TRUE) %>%
  # discard variables not being used
  select(activity, sum, day_of_week) %>%
  # widen table
  pivot_wider(names_from = "activity", values_from = "sum") %>%
  janitor::clean_names() %>%
  # replace the na values with zero (so we can calculate mean later)
  replace_na(list(stat_231 = 0)) %>%
  replace_na(list(stat_370 = 0)) %>%
  replace_na(list(amst_225 = 0)) %>%
  replace_na(list(cosc_283 = 0)) %>%
  # rename some variables
  rename("Mixed-Race America" = amst_225,
         "Theoretical Statistics" = stat_370,
         "Data Science" = stat_231,
         "Networks" = cosc_283) %>%
  # lengthen the table
  pivot_longer(cols = 3:6, names_to = "class", values_to = "time") %>%
  ungroup() %>%
  # calculate the mean of the time spent on each class by day of the week
  group_by(day_of_week, class) %>%
  summarize(mean_duration = mean(time))
```

```
# wrangle on the large dataset
activity <- calendar_data %>%
  # generate some summary statistics
  group_by(activity) %>%
  summarize(duration = sum(duration),
            count = n(),
            average_time = sum(duration)/n()) %>%
  # arrange in descending order by duration
  arrange(desc(duration)) %>%
  # give the classes more meaningful names
  mutate(Activity = case_when(activity == "amst-225" ~ "Mixed-Race America",
```

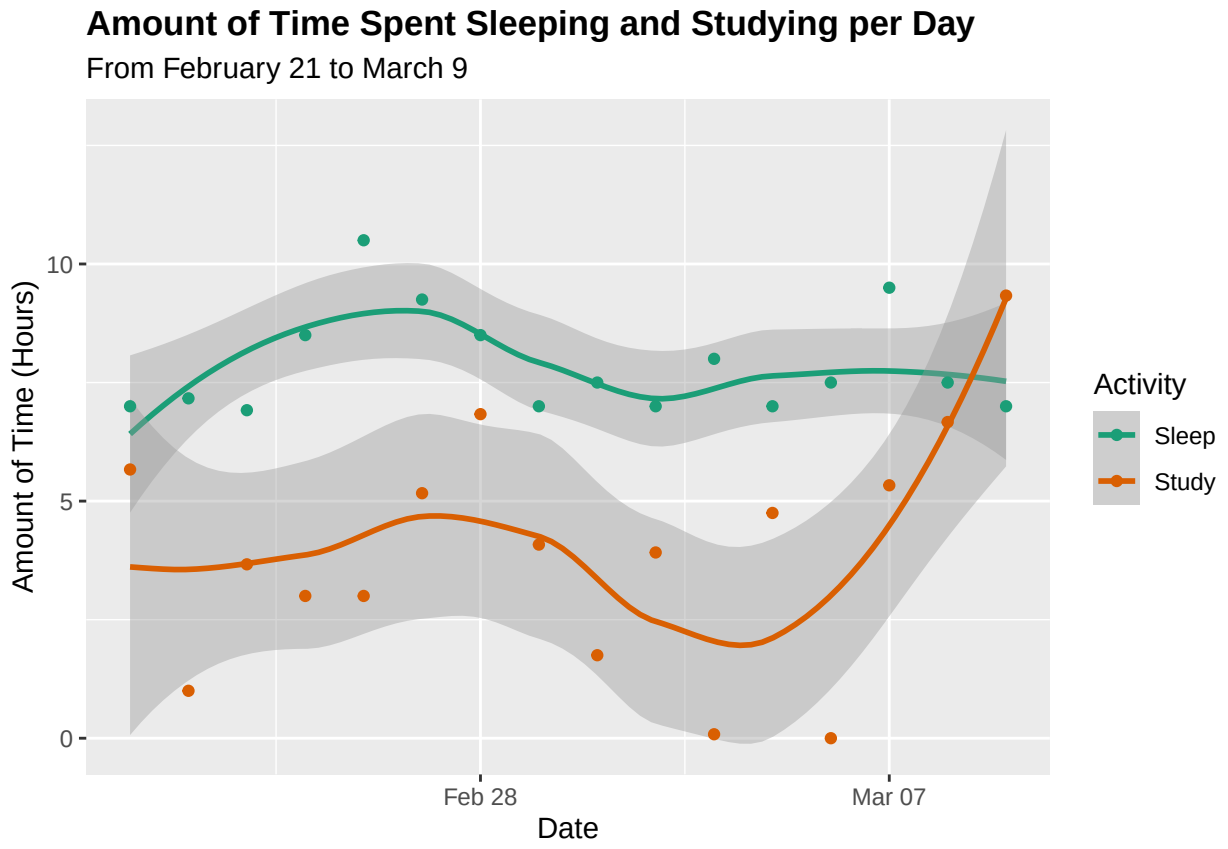
```
                              activity == "cosc-283" ~ "Networks",
                              activity == "stat-231" ~ "Data Science",
                              activity == "stat-370" ~ "Theoretical Statistics",
                              activity == "sleep" ~ "Sleep")) %>%
  # round some of the values and convert hours to minutes
  mutate("Total Time (in hours)" = round(duration,1),
         "Number of Sessions" = count,
         "Average Minutes per Session" = round(average_time*60, digits = 2)) %>%
  # remove unnecessary variables
  select(5:8)
```

Results

Visual 1

The first visual depicts the amount of time I spend sleeping and studying per day. Although there more data should be collected to determine any patterns, it appears that I spent more time sleeping than studying most days. The amount of time I slept appears to stay more consistent (less variable) throughout the time period while the amount of time I spent studying is more variable. Additionally, it appears that my peaks in sleeping also corresponded with peaks of studying. On average, if my sleep total is higher, my study total is also greater.

```
g <- ggplot(data = sleep_and_study, mapping = aes(x = date, y = time)) +
  geom_smooth(aes(color = activity)) +
  geom_point(aes(color = activity)) +
  labs(title = "Amount of Time Spent Sleeping and Studying per Day",
       subtitle = "From February 21 to March 9",
       x = "Date",
       y = "Amount of Time (Hours)",
       color = "Activity") +
  theme(plot.title = element_text(face = "bold")) +
  scale_color_brewer(palette = "Dark2")
g
```



Visual 2

The second visual depicts the average time I spend studying each day of the week. Each day of the week is further broken down into the average amount of time I spend studying for each class for that day. It appears that, on average, I spend the most time studying on Monday and the least amount of time studying on Friday. It appears that for each of my classes, I spend a greater amount of time on the coursework at various parts of the week. For example, I spend more time on Data Science work early on in the week (Monday, Tuesday, etc) while I spend more time on Theoretical Statistics later (Thursday). Additionally, on average, I spend more time on coursework during the beginning of the week (Sunday to Tuesday), and then the average amount of classwork I do, on average, decreases.

```
g <- ggplot(data = classes, mapping = aes(x = day_of_week, y = mean_duration, fill = class)) +
  geom_col() +
  scale_fill_brewer(palette = "Dark2") +
  labs(title = "Average Time Spent Studying by Day of Week",
       x = "Day of Week",
       y = "Amount of Time (Hours)",
       fill = "Class") +
  theme(plot.title = element_text(face = "bold"),
        legend.position = "bottom")
g
```

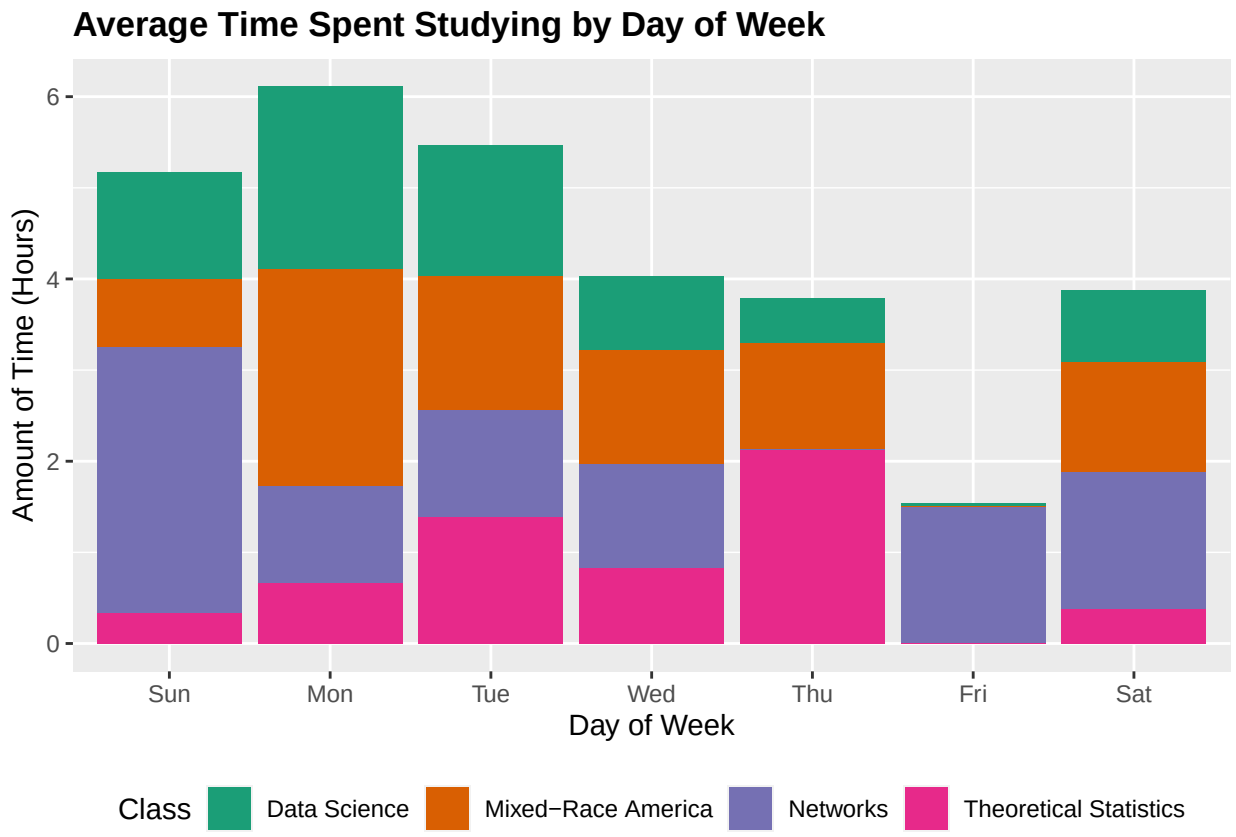


Table 1

The table below describes how much total time I spent doing each of my activities during my data collection window. Unsurprisingly, I spent the most time sleeping and the least time on Theoretical Statistics classwork. Of my four classes, I spent a similar amount of time on Mixed-Race America and Networks and a similar amount on Data Science and Theoretical Statistics. I tracked the most sessions for Mixed-Race America and the least for Networks of my classes. However, when looking at the average minutes per session, my sessions where I worked for Networks were the longest. It also shows that my average sleep time is 444.12 minutes (or about 7.4 hours).

```
activity %>%
  kable(booktabs = TRUE)
```

Activity	Total Time (in hours)	Number of Sessions	Average Minutes per Session
Sleep	125.8	17	444.12
Mixed-Race America	20.8	15	83.33
Networks	19.0	8	142.50
Data Science	16.6	11	90.45
Theoretical Statistics	14.0	13	64.62

Conclusions

After analyzing the data that I collected, I learned more about how I spent my time sleeping and studying. From my time plot, I was able to see specific trends over the roughly two weeks of my data window. There is not enough data to see any significant patterns, but on average, the amount of sleep I get on any given day is most likely greater than the amount of studying that I am doing. It is good to see that the amount of sleep I am getting is staying relatively consistent and hovering around 7.5 hours. This means that I am getting about as much sleep as I should be getting, which is key to having healthy sleeping habits. It is also interesting to see the time I am studying spike near the end of the time plot. This is most likely due to that week being "midterm week," where students have exams and larger projects due.

Through my bar plot, I was able to see, on average, how much time I spent studying for each class every day of the week. I was able to see that, on average, I spent more time studying during the beginning of the week and less time during the end of the week. I find it helpful to work ahead and get the most work done as early as possible, which could explain this pattern. My courses also meet on different days and have different assigned due dates throughout the week, which could explain why my average work time over the week differs from class to class.

Using my table, I was able to see that I got about 7.4 hours of sleep on average per day. Additionally, I spend the most time on my Mixed-Race America class and the least on Theoretical Statistics. Even though I spend the most time total on Mixed-Race America, on average, my sessions of Networks are the longest. One possible explanation for this is that my workload for Networks is more project-based, while Mixed-Race America has readings for homework. Readings usually take a shorter amount of time to complete than projects.

It is interesting to see the breakdown of how I spend my time, although collecting my own data could have introduced some bias into the analysis. Nevertheless, I am satisfied with the data that I have collected, so I will most likely continue to spend my time sleeping and studying the same as I did the past couple of weeks.

Reflection

Overall, I did not have many difficulties in the data collection process because it was relatively simple to input activities in a calendar on my phone or laptop, however as I mentioned above, it is difficult to tell how accurate the data is because we are essentially self-reporting the data. Because I knew that I was collecting my data for a specific time interval, some of my study habits unintentionally changed, such as switching between coursework less frequently or feeling more compelled to not any breaks. I attempted to recognize this bias, but there were not many options to work against it. It was also challenging to remember to record each activity that I was doing, and I would occasionally forget and try to remember the past day's events. In order to prevent that from happening, I tried recording the event as soon as I finished. A difficulty that I had in the analysis process was wrangling the data and finding summary statistics (such as means) when there were na values. I was unsure why my means were too large, and I had to redo much of the wrangling I had previously done.

For future data collection, it would be more accurate if we were not recording our data (maybe a computing device could) because it introduces bias when we know the experiment that is in progress. Most of the error in this project is the human error from not reporting an activity or inputting the wrong interval. If there was some way to do a blind experiment to collect the data, we could be sure that we have accurate data; otherwise, we would have difficulty finding any significant results.

There was not enough data for this project because it was difficult to see any trends or patterns for two weeks. In order to see trends, probably two months (or more) would be sufficient. This would account for any outlier weeks (such as Spring Break or Midterm Week) in which the data collected does not reflect regular habits. We would also see more trends in the time plot and more values to balance out the means. Because we only collected data for roughly two weeks, the means are very variable and can be shifted significantly by a single value. It would not be easy to collect this data because two months is a long time to be recording every time you do a particular activity.

As someone who provides data, I expect that my data is kept private and only used for the actions I authorize it to be used for. Additionally, I expect my data to be kept securely so that no one can steal my information. It is also expected that my data is not used in any unethical or harmful ways.

As someone who analyzes data, I have the ethical responsibility not to share the person's data without their consent or misuse it in any way. It is also my responsibility to recognize any biases associated with the data or the collection of data. As well as recognizing this bias, it is essential to act against letting bias influence models and analysis.