How I Spend My Time

SAT 231: Calendar Query

Kauila Tengan

Last updated October 8, 2021

Data Wrangling

```
# Data import and preliminary wrangling
calendar_data <- "ktdata3.ics" %>%
  # Use ical package to import into R
  ical_parse_df() %>%
  # Convert to "tibble" data frame format
  as_tibble() %>%
  mutate(
    # Use lubricate package to wrangle dates and times
    start_datetime = with_tz(start, tzone = "America/New_York"),
    end_datetime = with_tz(end, tzone = "America/New_York"),
    duration_min = difftime(end_datetime, start_datetime, units = "mins"),
    duration_hours = duration_min/60,
    # duration_min = end_datetime - start_datetime,
    date = floor_date(start_datetime, unit = "day"),
    # Convert calendar entry to all lowercase and rename
    activity = tolower(summary),
    overall = fct_collapse(factor(summary), #new overall variables for work/class
                           class = c("ASLC_class", "ASLC_study",
                                     "ENST_class", "ENST_OH", "ENST_study",
                                     "NS_class", "NS_study", "STAT_class",
                                     "STAT_OH", "STAT_SDS", "STAT_study"),
                           work = c("work_divtern", "work_meeting", "work_NISA")
                           ),
    #making duration into a numeric
    time = as.numeric(duration_min),
    time_hrs = as.numeric(duration_hours))



  #data wrangling for visualization 1
  class_data <- calendar_data %>%
     #filter our overall variable by class only
    filter(overall == "class") %>%
    #separating class and activity
    separate(summary, c("class", "activity"), "_", remove = FALSE) %>%
```

```
    group_by(class, activity) %>%
    summarize(overall_time = sum(time_hrs))


# Compute total duration of time for each day & activity
activities <- calendar_data %>%
  group_by(date, overall) %>%
  summarize(duration_min = sum(duration_min))
```

Introduction

Transitioning back to four in-person classes has been very overwhelming. Since I was off campus for the entirety of last year, returning here has made it difficult to balance outside life with school work. I feel as if I am doing too much and not enough. I often find myself saying, "I will do do that fun activity later ... when I have time for it." But then that time never comes. I was excited to participate in this calendar query project to investigate whether or not these feelings were supported by concrete data. Either way, I know that I will gain a lot of insight through this experience. I hope that this project will help me develop a better time management plan moving forward!

Questions and Design

Questions of Interest:

Question 1: *How much time do I spend on each course?*

For my first question, I wanted to explore how much time I was spending on each class. Within this question, I had two aspects of interest. Looking solely at the class data in my calendar, I wanted to see how total time spent on each class compared to one another. Secondly, I wanted to understand the breakdown of each individual class. Within a given class, I wanted to know how much time I devoted to studying, actually attending the class, visiting office hours, and if applicable, how much time was spent attending SDS hours.

Question 2: *How did my times spent on each activity during those two weeks compare to one another?*

For my second question, I wanted to look at the time allocated to all major activities during my day. I wanted to explore any interesting trends over the two weeks of data collection. I was curious how much of my day I was devoting to school, exercise, self care, leisure, sleep, and paid work.

Data Collection & Variables:

Going into the project, I already had an idea for my visualizations and table. I inputted all of my data into one Google calendar. I knew that I wanted to see both the breakdown within my courses, as well as the overall course data. I decided to stay consistent with my naming of each activity, paying particular attention to the capitalization or format of the letters. For each of my four classes, I abbreviated the department in all capital letters (ex. Data Science → STAT). I took it one step further for the specific activity. The capitalized abbreviation would then be followed by an underscore and then the activity in all lowercase (ex. studying for Data Science → STAT_study). When I began coding my visualizations, I realized that I only had one variable. Since my class and work inputs were highly specified, I had to first create a new variable `overall` and collapse my class and work data into `class` and `work` to address my second question. In order to create my stacked bar chart, segmented by activities, I had to then create an `activity` variable by separating `class` by the underscore. Here is a short breakdown of my variables and the data within each variable:

Visualization 1

`ASLC` : Media History of Anime Seminar

`STAT` : Data Science

`NS` : Culture and Mental Health

`ENST` : Environmental Studies Senior Seminar

`study` : Total time spent reading, completing problem sets, and writing papers

`OH` : Office hours

`SDS` : SDS Fellows office hours

Visualization 2

`class` : Aggregate time spent attending all classes, office hours, and studying

`gym` : Time spent at the gym

`rugby` : Time spent attending rugby practice

`sleep` : Estimated time I fall asleep until I wake up

`work` : Total time on paid work, attending club meetings, and any other meetings or work not related to academics

Results

Visualization 1: Breakdown of Total Time (Hrs) Per Class

Discussion:

The most noticeable thing about this visualization is just how little time was spent on my Culture and Mental Health class compared to the other three. This Hampshire College course is scheduled to meet on Tuesday and Thursday for an hour and twenty minutes each day. By absolute coincidence during these two weeks of data collection, both September 16th and September 23rd were cancelled for Yom Kippur and Hampshire Advising Day, respectively. This is reflected by the fact that the time in `class` and `study` is less than half of any of the other classes.

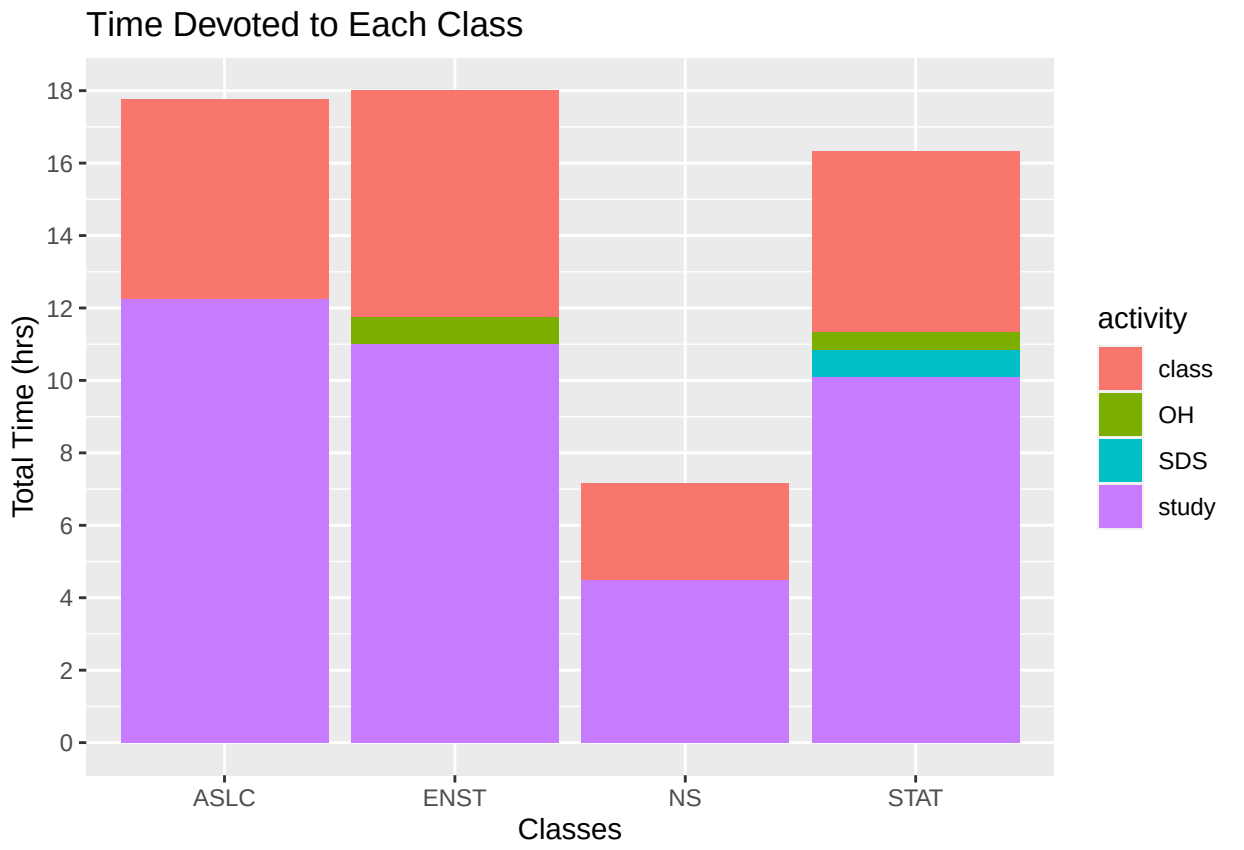
The distributions for my other three courses were similar and what I would have expected to observe. `ASLC` and `ENST` are both roughly three-hour seminars that meet once a week, so that class time should be greater than `STAT`. Additionally, this data shows me that I could, and should, be utilizing office hours more to reduce the overall time I spend studying. During those 16 days (384 hours), I only attended office and SDS hours a total of about two hours across those three classes. Contrary to the matter, I spent between 10 and 12 hours per class studying. Instead of struggling through problem sets, readings, and papers on my own, I would like to attend more office hours in the future.

Code and Output:

```
#creating plot from data in data frame class_data, segment the class by activity

ggplot(data = class_data, aes(fill = activity, y = overall_time, x = class)) +
```

```
#create a stacked segmented bar chart
geom_bar(position = "stack", stat = "identity") +
#clean up the labels, add title
labs(title = "Time Devoted to Each Class",
     x = "Classes",
     y = "Total Time (hrs)"
     ) +
#add more breaks on the y axis
scale_y_continuous(breaks = scales::pretty_breaks(n = 10))
```



Visualization 2: Time (Hrs) Devoted to Each Activity Per Day

Discussion:

Since I was tracking so many different activities, the graph is very cluttered and in particular regions, indiscernible. There are three major things that this graph describes. On average, I tend to get between seven and nine hours of sleep. However, on the night of September 20th, I had an overwhelming work load. I did work until three o'clock in the morning, but had a nine o'clock class at Hampshire College. Time spend on class (orange) increased from 3.5 hours the previous night to 5.5 hours. Since sleeping (purple) and being awake are mutually exclusive, my sleep significantly decreased from 9 hours to 4.5 hours.

The second major facet was `selfcare` and `leisure`. After the rigorous academic week, I made sure to spend the entire weekend healing. I walked around the art museum, took hikes in nature, journaled, meditated, and spent time with friends. My recorded self care on Saturday was 5.5 hours. I also noticed an interested day-by-day trend; when my class work increases, so does my leisure (green) time. I make a conscious effort

to balance my workload with leisure. If I do school work for a few hours, I then reward myself with time to watch my favorite shows or to play video games that same day.

The last thing that stood out to me was my inability to attend rugby (aqua) practice. Rugby is basically nonexistent on my graph. During the 16-day period, I was only able to attend two practices due to conflicts with other commitments. I was already considering quitting the team due to my lack of involvement, and this graph elucidated the fact that I simply may not have time for that particular activity.

Code and Output:

```
#create a time series separated by each overall activity
ggplot(calendar_data, aes(x = date, y = time_hrs,
                          color = overall)) +
  geom_line() +
  #make the labels look like and convey accurate information
  labs(title = "Time Series Graph Over 2 Weeks",
       subtitle = "Daily Allocation of Time to Each Activity",
       y = "Total Time (Hrs)",
       x = "Date (YYYY-MM-DD)",
       color = "Overall", lty = "Overall") +
  #Change date breaks on x axis so I can see every day over the two weeks
  scale_x_datetime(date_breaks = "1 days") +
  #Increase y labels (add 10 breaks)
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  #Rotates x axis labels 90 degrees and size for readability
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = rel(1)))
```

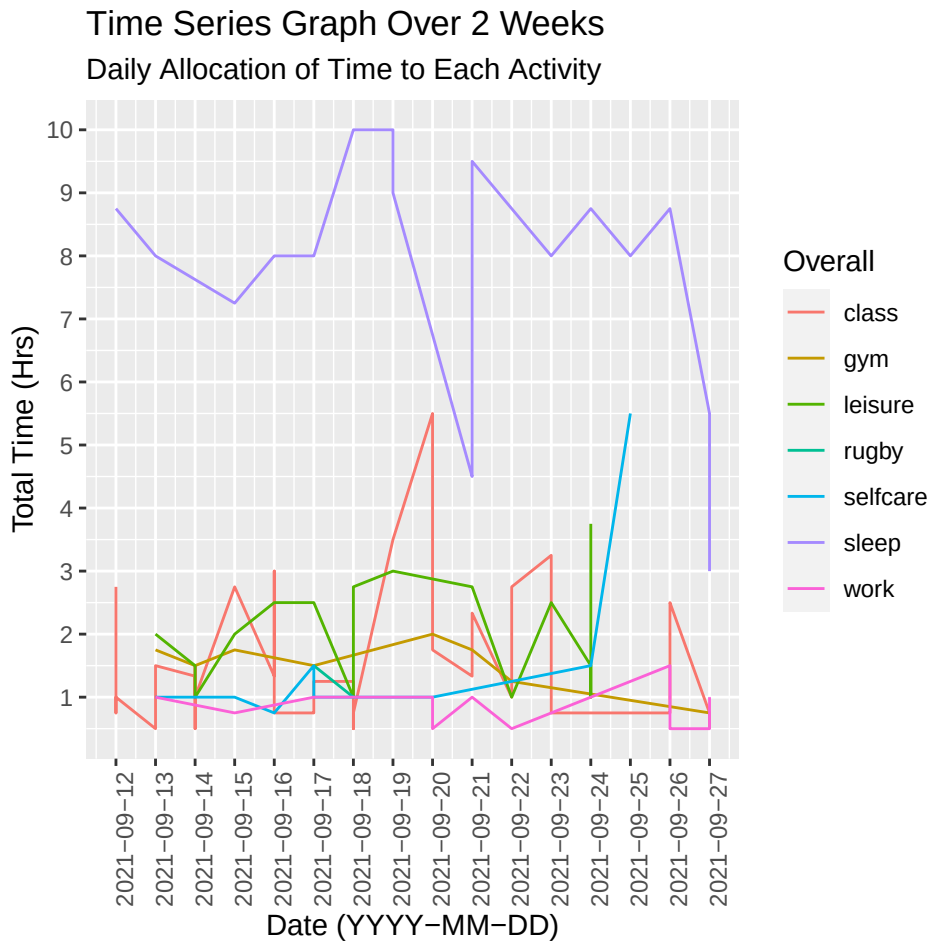


Table 1: Overall Summary of Each Activity

Discussion:

This table describes the average, minimum, maximum, and total hours devoted to each activity. `N` also indicates the number of times that activity was recorded in my calendar.

Due to the way I collected and wrangled my data, the average, minimum, and maximum hours fall short of conveying effective comparable data. These benchmarks are simply relative to the number of observations. For `sleep`, this table shows that whenever I inputted sleep data, it was for an average of 7.8 hours. On the other hand, `class` refers to any time input that was academic related. This includes times that I attended office hours for only 30 minutes, or if I had logged 40 minutes of work between classes. There were 37 recorded instances of anything school related.

In my opinion, the most important column in this table is `Total`. By far, I spent the most time sleeping recording a total of roughly 125 hours. This is merely 5 hours less than the total sum of the other six recorded activities (130.3 hours). Whenever I was awake, most of my time was spent attending classes and doing homework. Total `class` time was about 60 hours. I also participated in leisurely activity a lot when compared to other facets of my life. Excluding sleep, whenever I participated in a given activity, I devoted an average of about 1.5 hours to it. An important thing to note is that the number of observations differed significantly. Average hours for both class and gym were about 1.5 hours but I did school-related activities 30 more times than I went to the gym.

Table 1: Activity Summary Statistics (Hrs)

Activity	N	Average	Min	Max	Total
sleep	16	7.8	3.0	10.0	125.0
leisure	16	2.0	1.0	3.8	32.2
class	37	1.6	0.5	5.5	59.2
selfcare	9	1.6	0.8	5.5	14.2
gym	8	1.5	0.8	2.0	12.2
rugby	2	1.2	1.0	1.5	2.5
work	12	0.8	0.5	1.5	10.0

Code and Output:

```
# Create new data frame from data frame calendar_data
table <- calendar_data %>%
  # I am interested just in the overall data
  group_by(overall) %>%
  #Renaming variable overall
  rename(Activity = overall) %>%
  summarize(
    N = n(),
    #Create an average column
    Average = mean(time_hrs),
    #Create a min time column
    Min = min(time_hrs),
    #Max time
    Max = max(time_hrs),
    #Total time
    Total = sum(time_hrs)
    ) %>%
  #arrange in descending order
  arrange(desc(Average)) %>%
  #make aesthetically pleasing and change number of digits following decimal.
  kable(booktabs = TRUE, digits = 1, caption = "Activity Summary Statistics (Hrs)") %>%
  #center and change font size
  kable_styling(font_size = 15)

#output table
table
```

Conclusion

In summary, the amount of time I devoted studying and attending classes were relatively even. The only outlier was my Culture and Mental Health Class, but the lack of time was due to the circumstances of the sampling period. It was clear that I spent a lot of time studying by myself and hardly utilizing office hour support. Moving forward I hope I will make a conscious effort to attend office hours when I feel overwhelmed in hopes that it will reduce my overall study time. Analyzing the time series graph and the `Average` column of the table, I have been balancing my work well with other activities. Whenever I participated in an activity (excluding sleep), I devoted on average about 1 - 1.5 hours to it. Looking at the observations (`N`), I was only able to attend two rugby practices over two weeks. One factor possibly contributing to my being overwhelmed is that I have been doing a lot of things. I now see that I can utilize office hours better and perhaps slightly reduce sleep and leisure to get more work done. Although this testing period is merely a parcel of my semester, I feel like I have learned a lot.

Ethical Reflection

The first hurdle was recording everything under one variable. In order to produce my visualizations, I had to do a lot of data wrangling to set up my data frames appropriately. The second hurdle was that I had misspelled some of the class abbreviations. I only realized this after producing my bar chart for the first time. Because of this, I had to revisit my calendar, word search the misspelled names, and then re-export the calendar data.

For future data collection, it is always important to consider reproducibility of my work. How can I can create a data set that can be easily revised or edited with additional data in the future? The major thing for me is consistency, especially when working with coding structures extremely attentive to capitalization or characters (i.e. "STAT_activity"). If one character is misspelled, it will result in a separate category of a bar chart like it did for mine. One of my tasks this summer was updating statistics on a research paper for an environmental consulting firm. The additional 5 years of data were taken differently and the cells in excel were not all the same format, some quantitative columns were numerical values and others were strings saying "120 ml." The other major issue was that they had none of the code to produce any of the graphs again, so I had to go in and essentially start all over from scratch. Small details in the way data is inputted (i. e. capitalization or type of cell) can ultimately be the difference of a few minutes or a few days of work for a data scientist. This was my experience this summer and something that I was particularly attentive to going into this project. With that said, I also made sure to comment each of my code lines for reproducibility in the future.

In order to fully answer my questions, I think I would have to collect data for an entire year. Since one of my classes was cancelled two weeks in a row, the data portrayed a very inaccurate image of how my time is really spent. If I included more observations for `NS` I am confident that the bar chart would look very similar to my other three classes. I would also be interested in how other variables affect my activities and time commitments. I would hypothesize that time spent studying and sleeping would drastically shift during midterm/final exam periods. As far as other activities, I most likely will not be spending as much time outside exercising, playing rugby, and walking in the forest during the dead of winter with negative temperatures and ice everywhere. Although I need a lot of data to form better conclusions, that also requires an incredible amount of time and diligence. This is often the major limitation in all scientific peer-reviewed journals - the lack of time and money for long-term studies.

Lately I have been spending a lot of time considering the ethics of data science. In my Environmental Studies Seminar, we have been learning how peer-reviewed journals, funded by pesticide companies straight up lie about their statistics. They withhold information about the design of the experiment and when other data scientists attempt to reproduce these findings, it is impossible because the findings are fabricated. This data then goes to the EPA who then permit the use of these carcinogenic substances on all of our food. These ethical quandaries are prevalent in chapter 8 of our textbook. In this circumstance, data science in weaponized to allow for the displacement and potential killing of millions of people. Another major issue is privacy and plagiarism. As for the Big Tech question, the obvious expectation is that they do not sell my personal data. However, this is just the reality of today. Companies will sell information for revenue from advertisement companies. When I think about this data, I am scared that the demographics may be used to legitimize and reproduce aspects of scientific racism. To be honest, I do not have a clue what will happen, but I am very cynical about knowledge in the hands of the wrong people. As someone who analyzes others' data, especially when scraping, I must make sure that paths are allowed to avoid plagiarism or deleterious affects to their servers. There are also aspects of confidentiality. If someone does not want to be named or associated with data that may affect them, I must respect this.