Distribution of Time Among Pairs of Activities

STAT 231: Calendar Query

Cassandra Jin

Last updated March 12, 2022

Introduction

My questions of interest are the following:

- Do I spend more time practicing violin or bass guitar?
- How much of my total sleep time do my sporadic naps take up?
- Do I make efficient use of my entire week for completing my math assignments, i.e. do I spread out my workload for the classes over the course of the week?

I've been playing violin for most of my life but decided this semester to start taking bass guitar lessons. I certainly enjoy learning bass, but taking lessons for it adds a faint aspect of obligatory technical progress on the instrument, so I sometimes feel less inclined to pick it up and fiddle around on it for fun. So, I am interested to see whether I gravitate toward practicing for bass (out of necessity for the lessons) or practicing for violin (mostly for fun). As for my sleep schedule, I have very poor time management skills and napped for more time than I slept last semester. I am making an effort to change that and would like a quantitative visual of my current sleep schedule. In lieu of this, my academic work schedule is just as poorly organized. Every assignment requires that I put in a certain minimum of hours, but I always start a day or two before it is due and end up working for very long bouts of time in few sittings (late at night, too). Perhaps seeing the answer to my third question will show me that spreading out my workload over time can give me the same result for less stress.

Data collection

Since I had relatively simple tasks to record, I collected data by noting the time chunks I had spent after finishing an activity. My only quantitative variable is time (in minutes), while my categorical variables span my six activities, which I will compare in pairs (bass v. violin, napped v. sleep, 260 v. 355).

```
# Data import and preliminary wrangling
calendar_data <- "project1_data_03_10.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
    ## Feel free to use minutes instead
    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(start_datetime, unit = "day"),
    ## Examples of ways to parse dates, times (keep only what you need!)
    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)
```

There will be three plots and one table. For my first question regarding instrument practice, I plan to use a grouped bar chart involving coloring by activity so as to compare the duration of time spent on violin and that on bass over time. My second plot will be a stacked bar chart that allows the reader to see the amount of time that I nap and that I sleep at night as portions that make up the total amount of sleep I get in a day. For this same question, I will present a table containing the five-number summary of my sleep throughout the week. For my third question about math workload distribution, I will use a bar chart with a smoothing function to show the intervals of time in a day during which I choose to work on my problem sets.

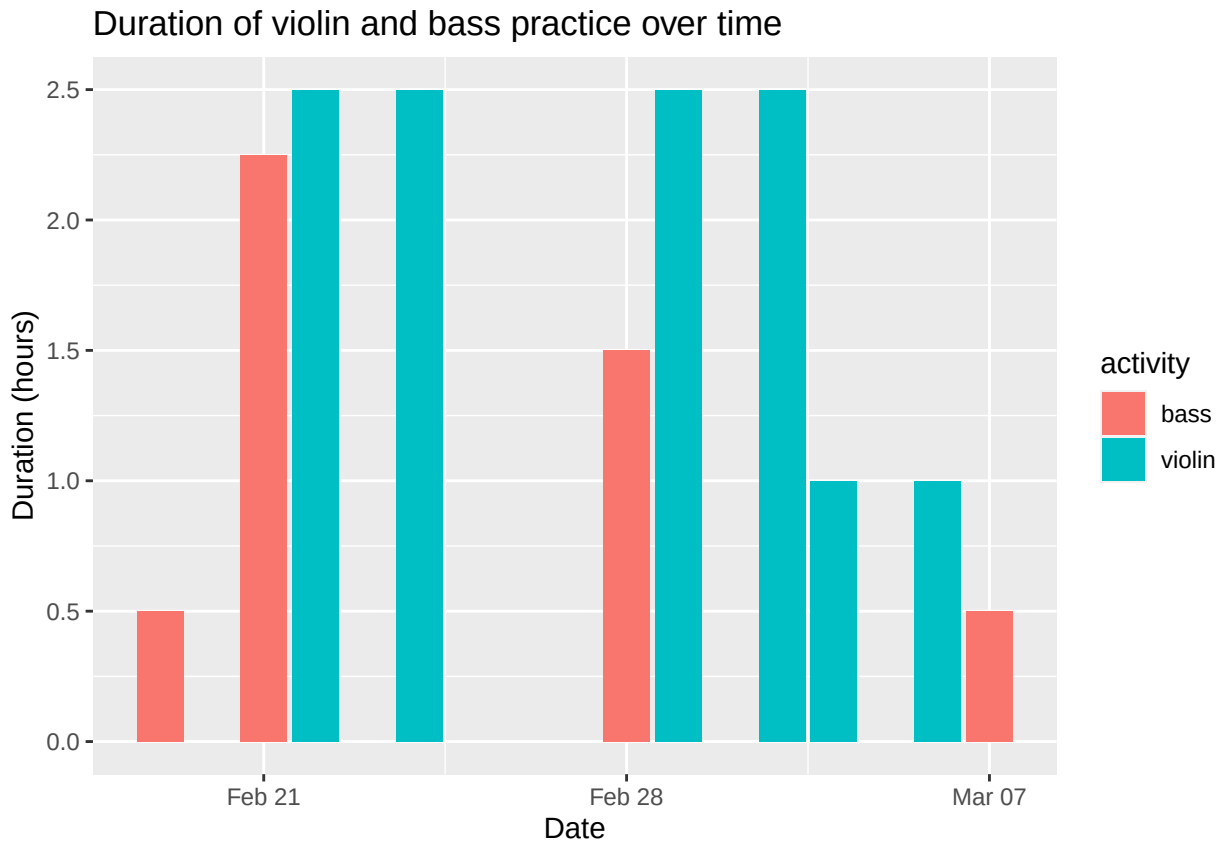
```
# Example of preparing dataset for first visualization

# Compute total duration for each activity and number of each
activities_total <- calendar_data %>%
  group_by(activity) %>%
  summarize(duration = sum(duration),
            count = n())
```

Results

The first data visualization is a bar graph that captures the relationship between time and the amount of time I spent practicing violin and bass. Time is linear (the entire data collection period rather than, say, day of the week), and the bars are colored according to either activity. At first glance, it seems that I practice violin more because of the bars reaching 2.5 hours. However, these are Amherst Symphony Orchestra rehearsals, so although I am using my instrument and improving on some pieces, I do not think of these as purely personal practice. There is a bar of bass practice at the beginning of every week (Feb 21, Feb 28, Mar 07), because I have my bass lesson then, so in preparation, I practice for a while the night before or the morning of. Removing, then, the four violin bars with duration of 2.5 hours, it appears that I spent more time practicing bass.

```
practice_data <- calendar_data %>%
  filter(activity %in% c("bass", "violin"))
```

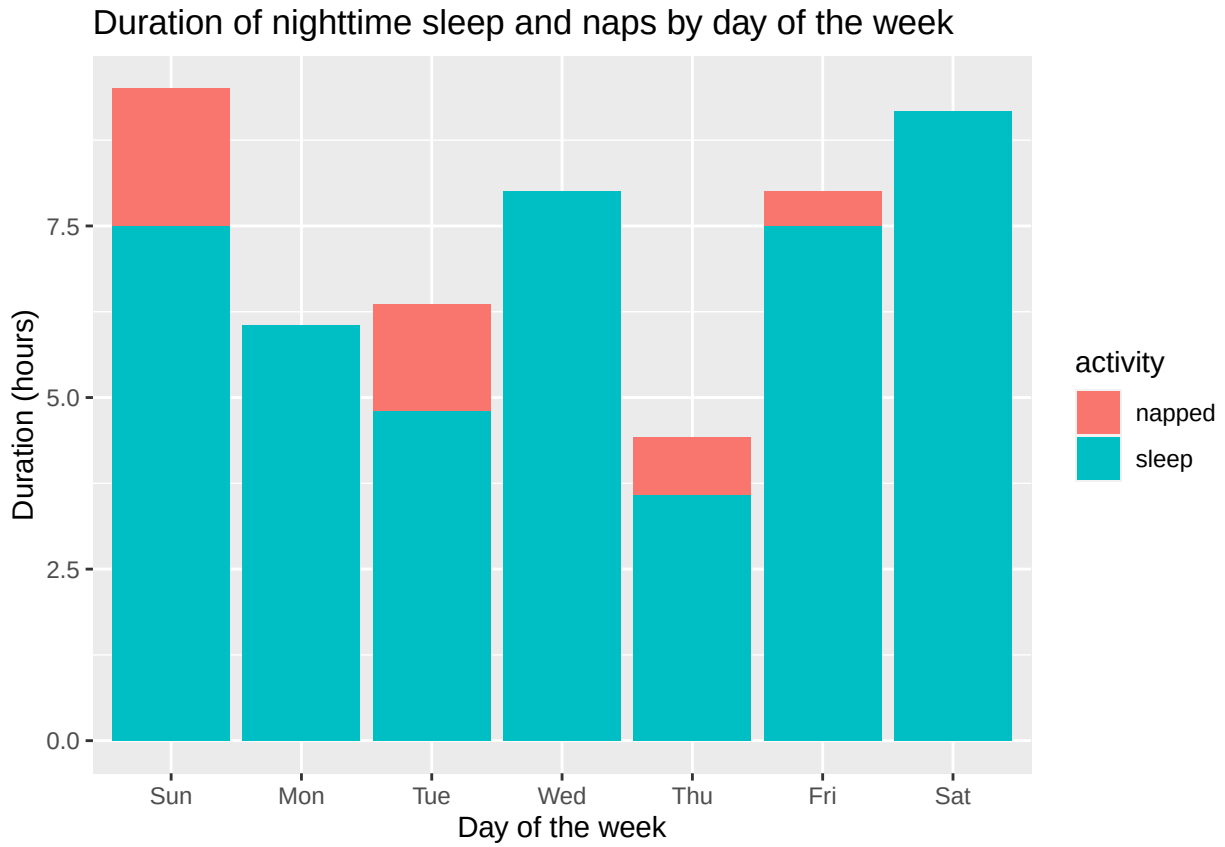
```
ggplot(practice_data, aes(fill = activity, y = duration, x = date)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(y = "Duration (hours)",
       x = "Date",
       title = "Duration of violin and bass practice over time")
```



The second data visualization is a stacked bar chart that shows the reader the durations of time for which I sleep at night (in blue) and nap (in pink) as portions that make up the total amount of sleep I get in a day of the week. Since the pink partial bars do not make up that much of my total sleep time and are not even present on some days, it appears that naps do not amount to that much of my sleep. In addition to the comparison between night and nap sleep, the graph contains decreasing bar heights as the days of the week progress. This plummet in duration between the beginning and late half of the week indicates that I sleep less and less, as assignments tend to be due at the end of each week.

```
sleep_data <- calendar_data %>%
  filter(activity %in% c("sleep", "napped")) %>%
  select(activity, day_of_week, duration) %>%
  group_by(activity, day_of_week) %>%
  summarize(mean_duration = mean(duration))
```

```
#ggplot(sleep_data, aes(fill = activity, y = duration, x = day_of_week)) +
ggplot(sleep_data, aes(fill = activity, y = mean_duration, x = day_of_week)) +
  geom_bar(position="stack", stat="identity") +
  labs(y = "Duration (hours)",
       x = "Day of the week",
       title = "Duration of nighttime sleep and naps by day of the week")
```



The following table contains the five-number summary of my total sleep time and conveys that it is skewed left, meaning that I have a few values of very large durations of sleep, but most of the recorded durations are the shorter side. I also included the mean of my total sleep values, which is less than the median and further supports that most of my data points lie on the lower end.

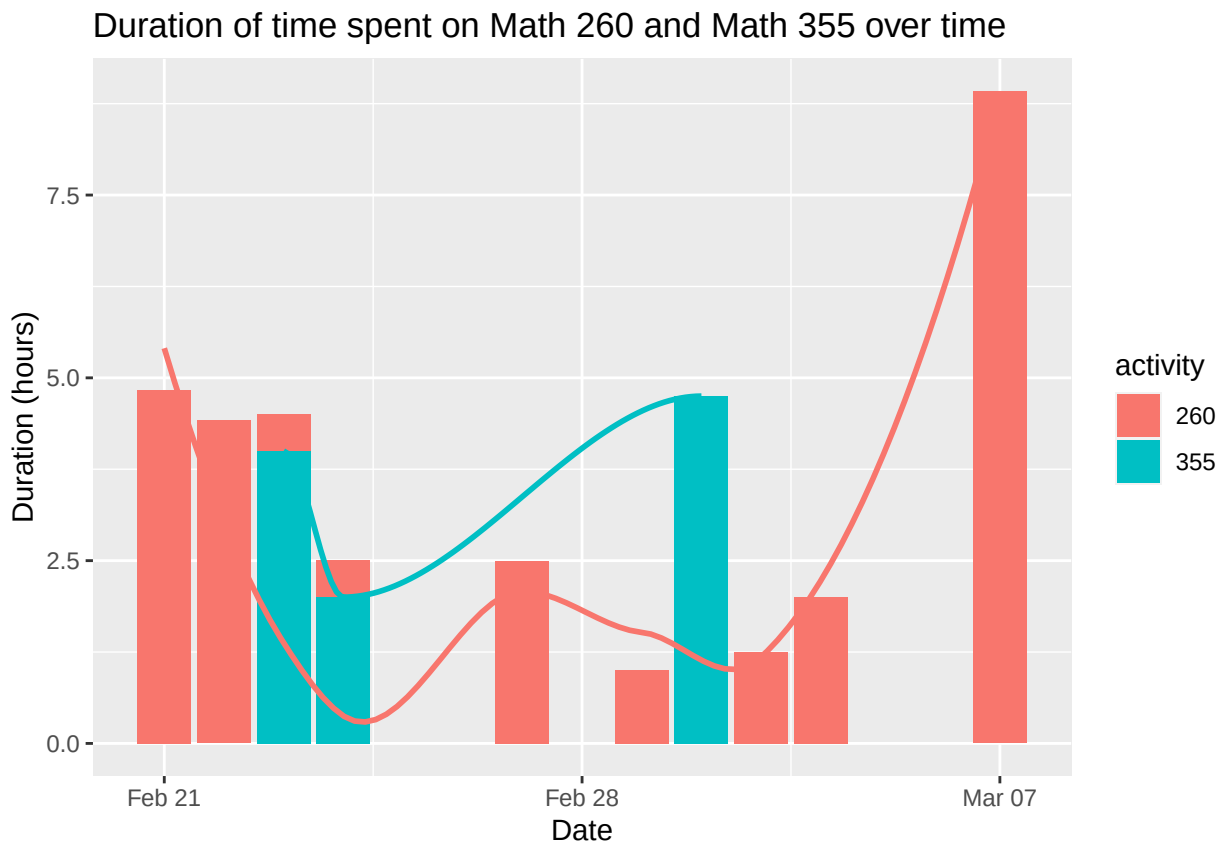
```
table_data <- calendar_data %>%
  filter(activity %in% "sleep") %>%
  select(duration) %>%
  summarise(min(duration),
            quantile(duration, 0.25),
            median(duration),
            mean(duration),
            quantile(duration, 0.75),
            max(duration)) %>%
  rename (Min = "min(duration)",
          Q1 = "quantile(duration, 0.25)",
          Median = "median(duration)",
          Mean = "mean(duration)",
          Q3 = "quantile(duration, 0.75)",
          Max = "max(duration)")

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

Min	Q1	Median	Mean	Q3	Max
2.666667	5	7	6.708333	8.375	10

The third visualization is a stacked bar chart that also uses the function geom_smooth(). It shows that, in some ways, I distribute the time that I spending working on Math-260 better than I do for Math-355, because the pink 260 bars are more frequent and similar in height than the blue 355 ones. Despite this, there is very obviously a wave shape to both smoothed lines, indicating that over the course of the data collection period of about 2.5 weeks, I clumped most of my work time into certain days of each week (before the due dates). Also the peaks of the two smoothed lines are slightly shifted from one another, with the 355 peak coming after the 260 one. This reflects the later weekly due date of the 355 problem set.

```
math_data_smooth <- calendar_data %>%
  filter(activity %in% c("355", "260")) %>%
  select(activity, date, duration) %>%
  group_by(activity, date) %>%
  summarize(duration_sum = sum(duration))
```

```
ggplot(data = math_data_smooth, aes(y = duration_sum, x = date, fill = activity)) +
  geom_col() +
  #geom_point() +
  geom_smooth(aes(color = activity),
              se = FALSE) +
  labs(y = "Duration (hours)",
       x = "Date",
       title = "Duration of time spent on Math 260 and Math 355 over time")
```



Conclusions

For the instrument question where I investigated whether I spend more time practicing violin or bass, I learned that I practice for bass most often only once a week, right before my lesson. This is a similar message to that of my third visual (math workload distribution over time). Like my approach to practicing for my bass lessons, I leave my work and studying to a day or two before the due dates. The second visual documenting my sleep durations, however, tells a more positive story. This is because I hope to nap less these days rather than sleeping very little at night and trying to make it up during the day. Since the bars of total sleep do not consist much of pink (indicating nap), I am improving in this aspect of relying less on naps. Going forward, I will try to do a little bit of work every day for every subject. Ideally, my first and third visualizations would contain bars of short but nearly equal height over time, meaning that I would be appropriately spreading out my practice and workload throughout the week. Also, I spend most of my time doing schoolwork, attending office hours and club meetings, having fun with friends, and trying to catch up on sleep. The first bar plot has shown me that I should allocate some time in my schedule to playing violin for fun, and perhaps this would contribute even more positively to my well-being than a nap does.

Reflection

In the beginning of the data collection process, I had slight difficulty in staying consistent with how I recorded data, including how specific the time ranges were and the names of the activities. Also, I am easily distracted while doing work during the time that I allocate to it, so I had to consider whether I wanted to record a duration as the time that I should have spent doing it or the time I actually did spend on it. In the analysis process, it took a while for me to understand what I wanted in my visualizations and how I should go about producing them. That is, I knew the questions that I wanted to investigate, but I often had to draw out a sketch of a plot, containing the axes, any visual cues, and a rough prediction of what the actual data would look like. Then figuring out how to wrangle and present the data became easier.

If I were to repeat this project, I would likely need to develop a note system to keep myself consistent when recording and analyzing data. Also, in addition to more concrete, quantifiable research questions, I would like to have a clearer goal for my visualizations. This would allow me to focus on making some aspects of my data collection more detailed while relaxing on others, which would give me more appropriately focused data to wrangle, and I would then perhaps be able to produce visualizations with less wrangling.

For my second and third questions involving sleep and workload patterns, I feel that I did not have enough data because certain trends that I had predicted began to emerge in the visualizations, but the data collection period was too short. I think that any inconsistencies in my schedule were more present as a result, so having many more weeks of data, maybe around three months, would yield a more accurate understanding of my normal schedule. The data that my questions require are not particularly hard to collect. The most difficult aspects of data collection would likely be simply remembering to collect it and staying consistent with how I record.

When I provide my data, I expect my anonymity and that my data will be used responsibly and for the potential good of others. I would also hope that such organizations would not use some components of my data as a proxy for others, because I've learned that a large amount of bias can stem from this and I do not wish to be used to perpetuate historical stereotypes into the future.

As someone who analyzes others' data, I would need to preserve their identities both from myself and from the public when the data are published (which is also a necessary step). I also have the responsibility to frame the results of the data in an understandable and entirely truthful manner so that not only am I performing data collection and analysis correctly, but the public is receiving the correct message. Among all of these steps, I need to ensure minimal or no bias, which is nearly impossible to achieve in many investigations but must always be kept in mind.