Work, Sleep, Repeat

Stat231: Google Calendar Report

Brandon Kwon

Due Friday, March 19 by 5:00 PM EST

How do I spend my time?

```
# include your import of the data and preliminary wrangling here
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ical)
library(ggcorrplot)
library(dplyr)
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
my_calendar0 <- ical_parse_df(file = paste0("~/Spring 2021/Data Science - STAT 231/bkwon23@amherst.edu.
  mutate(start_datetime = with_tz(start, tzone = "America/New_York")
    , end_datetime = with_tz(end, tzone = "America/New_York")
    , length_hour = end_datetime - start_datetime
    , date = floor_date(start_datetime, unit = "day")
    , day = weekdays(date))


# Obtain Necessary Columns
my_calendar1 <- my_calendar0 %>%
  select(c(summary, start, end, length_hour, date, day))

# Rename Columns & Filter to Designated Two Weeks
my_calendar1 <- my_calendar1 %>%
  rename(Activity = summary, Duration = length_hour, Date = date, Start = start, End = end, Day = day)

my_calendar1 <- subset(my_calendar1, Start > as.POSIXct('2021-02-28') & End < as.POSIXct('2021-03-14'))

my_calendar1 <- my_calendar1[order(my_calendar1$Start),]

my_calendar1$Duration <- round(my_calendar1$Duration/3600, 2)
attr(my_calendar1$Duration, "units") <- "hours"
```

Describe your question(s) here. Briefly describe your data collection process, including how you defined variables of interest. - What are the top ten activities that I commit to the most during these two weeks? - How do these top ten activities correlate with each other in terms of time?

I spent my time collecting data from the dates of 02/28 to 03/14. Because I was not set upon what types of questions I wanted to answer, I made sure to collect data on practically everything I was doing. In order to keep consistency and make data wrangling easier on my end, I kept names very consistent if the event was repeated (such as "CHEM 231," "Study/Homework," etc.). I also made sure to leave almost no gaps between my events so that I could have as much to work with as possible.

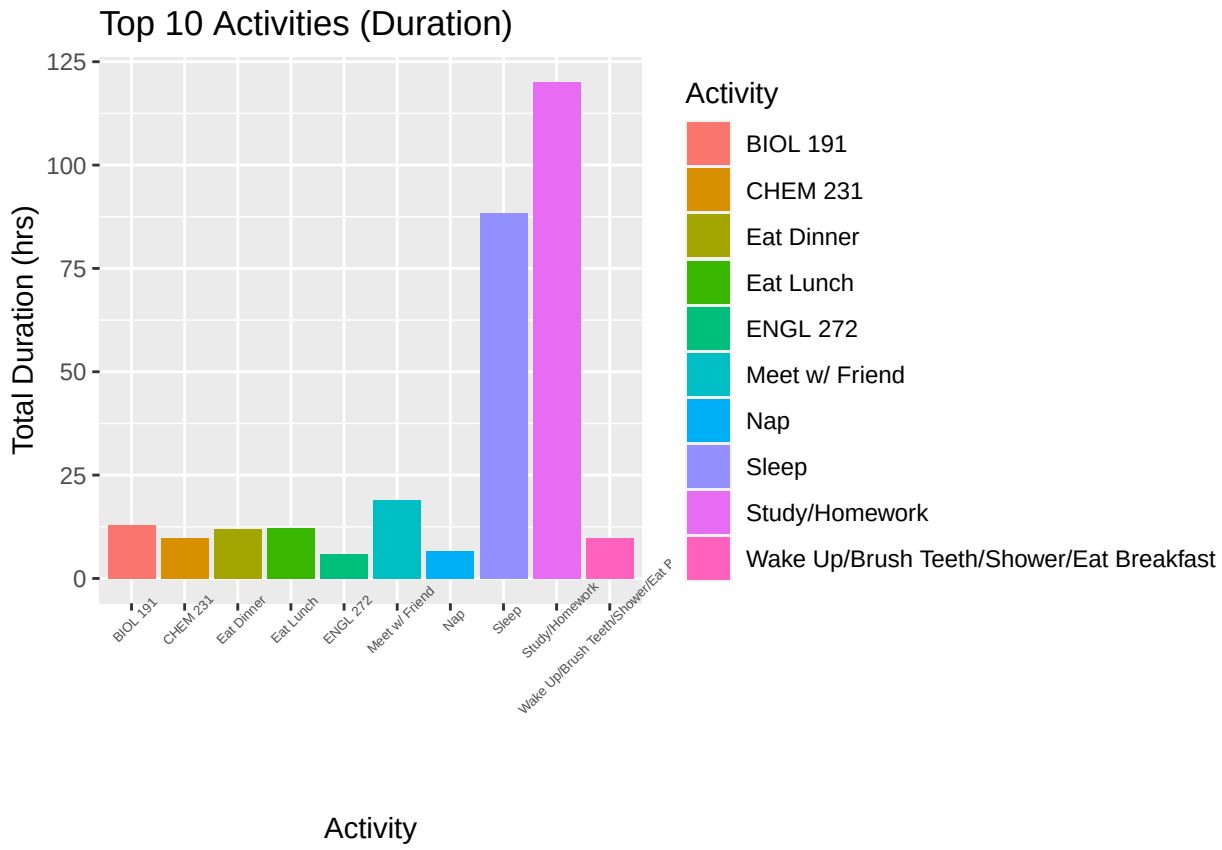
In terms of defining my variables of interest, I created names for each event that would be easier to recognize. Because most of the events that I accounted for wouldn't really need to be grouped by a certain category, I didn't worry too much about strict similarity of names besides keeping the same event consistent. Because my questions above tell an individualized story of each event rather than a story that groups the events with one another, I knew that as long as I had my data recorded, I could proceed with further analysis, as demonstrated below.

Describe what information is conveyed through data visualization #1 (that you'll create below) here.

```
# Write your code to create data visualization #1 here.
# Be sure to label your axes and include a title to give your data context.

# Aggregate Variables for Better Analysis
my_barcalendar1 <- aggregate(my_calendar1$Duration, by = list(my_calendar1$Activity), FUN = sum)

# Sort Aggregate Variables
my_barcalendar1 <- my_barcalendar1 %>%
  arrange(desc(x)) %>%
```

```
  rename(Activity = Group.1, Duration = x)


my_barcalendar1 <- head(my_barcalendar1, 10)

# Create Data Graphic (Bar Chart)
# Represents the total duration of the top 10 activities that I spent the most time on
ggplot(data = my_barcalendar1, aes(x = Activity, y = Duration, fill = Activity)) +
  geom_bar(stat = "identity") + xlab("Activity") +
  ylab("Total Duration (hrs)") +
  ggtitle("Top 10 Activities (Duration)") + theme(axis.text.x = element_text(angle = 45, size = 5))
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

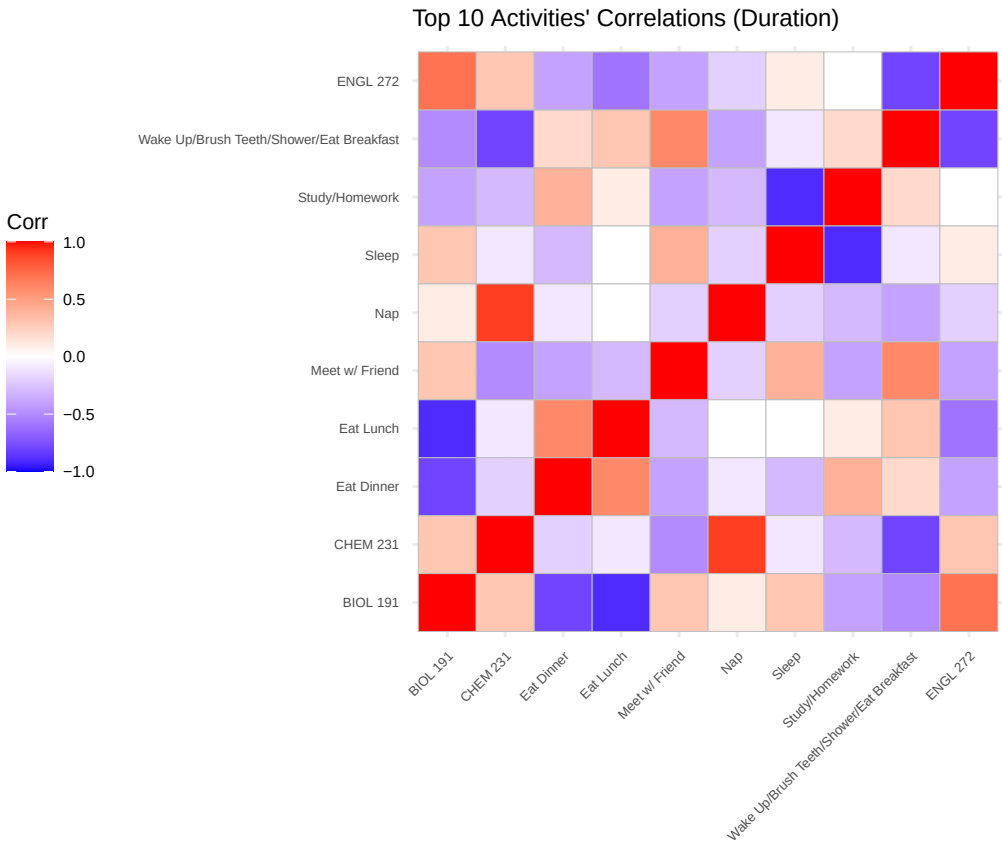


Describe what information is conveyed through data visualization #2 (that you'll create below) here.

```
# Write your code to create data visualization #2 here.
# Be sure to label your axes and include a title to give your data context.

# Create a subset of "my_calendar1" to keep track of duration of each activity in accordance to day
x <- filter(my_calendar1, Activity %in% my_barcalendar1$Activity) %>%
  select(c("Activity", "Duration", "Day")) %>%
  group_by(Day, Activity) %>%
  summarise(total_time = as.numeric(sum(Duration)))
```

```
## `summarise()` has grouped output by 'Day'. You can override using the `.groups` argument.
```

```
# Reshape the dataset in a way that widens the data set for better analysis
y <- x %>%
  pivot_wider(names_from = Activity, values_from = total_time, values_fill = 0)

# Get rid of "Day" column
 temp <- y[,-1]

# Create Correlation Matrix
corr <- round(cor(temp), 1)

colnames(corr) <- colnames(temp)

rownames(corr) <- colnames(temp)

ggcorrplot(corr, title = "Top 10 Activities' Correlations (Duration)", tl.cex = 5) +
  theme(text = element_text(size = 7.5), legend.position = "left")
```



Describe what information is conveyed through the table (that you'll create below) here.

```
# write your code to create the table here
# if you want to make your table's appearance nicer, check out:
# the xtable package (https://cran.r-project.org/web/packages/xtable/vignettes/xtableGallery.pdf); or
# the kable function in the knitr package (https://bookdown.org/yihui/rmarkdown-cookbook/kable.html)
```

Table 1: Table of Duration of Events by Day (Hours)

Day	BIOL 191	CHEM 231	Eat Dinner	Eat Lunch	Meet w/ Friend	Nap	Sleep	Study/Homework	Wake Up/Brush Teeth/Shower/Eat Breakfast	ENGL 272
Friday	3.66	1.66	1	1.00	7	2.00	12.00	15.50	1.66	0.00
Monday	2.50	3.99	1	2.00	0	2.50	13.33	13.83	0.66	1.33
Saturday	2.00	0.00	1	2.00	7	0.00	18.00	11.50	2.00	0.00
Sunday	0.00	0.00	2	2.00	2	0.00	7.00	26.00	2.00	0.00
Thursday	0.00	2.50	3	2.17	0	2.17	10.50	18.50	1.33	0.00
Tuesday	0.00	0.00	3	2.00	2	0.00	14.50	16.83	1.50	0.33
Wednesday	4.83	1.66	1	1.00	1	0.00	13.00	18.00	0.66	4.16

```
# Create table based on Top 10 Activities that I spent the most time on
knitr::kable(y, caption = "Table of Duration of Events by Day (Hours)") %>%
  kable_styling(font_size = 4)
```

To conclude, briefly summarize what you found in response to the questions posed here.

Based on my data graphics, I can see that I spent a disproportionate amount of time studying and doing homework (120.16 hours to be exact). While on the surface this seems pretty remarkable as a student, I must realize that there is more to life than just focusing on my academics. After conducting this project, I am more motivated to spending more time with family and friends as well as doing more outdoor activities. Based on my bar graph, I see that I spend an adequate amount of time sleeping as well, which should be normal considering that I need sleep to function well every day. To conclude, I can see that I spent the most time during these two weeks studying/doing homework, sleeping, meeting with friends, eating dinner, eating lunch, studying for BIOL 191, studying for CHEM 231, studying for ENGL 272, napping, and waking up/getting ready for the day.

A relationship that I found quite unsurprising when analyzing my correlation matrix includes the relationship between the amount of time I sleep and the amount of time I spend studying and doing homework. Based on the matrix, I see that the correlation between these two recurring events is one that is negative, almost to the point in which I can conclude that the correlation is STRONGLY negative. I can affirm this relationship because sometimes when I have a difficult work load on a certain day, I am more than sure that I will receive less sleep that day, mostly because I'll probably be up studying late at night (possibly until 2:00AM-3:00AM). Another relationship I found quite intriguing is one that involved napping and spending time on CHEM 231. For this positively correlated relationship, I cannot seem to address a reason for this phenomenon. As a result, I must be aware that correlation does not equate to causation (as I learned in STAT 230). This assertion can also definitely be supported by the positive correlation between my time working on BIOL 191 and my time working on ENGL 272.

In the grand of scheme of things, I've realized that while I was mostly productive during these two weeks, I found that my life can be quite lacking in terms of diverse activities. I hope to explore new ideas and commit to more fun activities that inspire me in the future. I hope that while school is very important, I focus on mental and social health as well.

Reflection

Write your one-page reflection here in paragraph form. In particular, address:

- What difficulties in the data collection and analysis process did you encounter? Identify two of your main hurdles in gathering accurate data.
- What implications does that have for future data collection and/or analysis projects?
- How much data do you think you'd need to collect in order to answer your question(s) of interest? Would it be hard to collect that data? Why or why not?
- As someone who provides data, what expectations do you have when you give your data (e.g. to Facebook, Google, MapMyRun, etc.)?

- As someone who analyzes others' data, what ethical responsibilities do you have?

When starting off this project by collecting data, I found that repeating events (such as classes) had to be accompanied by the same exact name. This made analysis much easier to complete. However, as someone who uses Google Calendar for just classes, I found that one main hurdle I had to overcome was recording everything I completed throughout these two weeks, including when I ate, when I worked out, and when I attended a meeting. Another main hurdle I had to leap over was estimating when I completed various events, mostly because whenever I finished a specific event (not classes), I would forget to immediately record it on my Google Calendar. Fortunately, I believe my estimations were as accurate as they could have been. After collecting my data, I found that coming up with creative questions that reflected my calendar was somewhat difficult to complete. There were so many questions I wanted to ask, but by making my questions more generic, I knew that I could elaborate more upon my analyses. As a result, I arrived at my designated questions, and luckily, my questions were related to each other in many ways. In terms of analysis, I found that creating the correlation matrix was somewhat difficult because I had to utilize Google as my main resource. What was difficult was not the actual creation of the matrix but rather the accurate tidying/creating of the sub-datasets necessary in order for them to be incorporated into the data graphic itself. To be more general, I also found choosing a data visualization graphic to be difficult because I wanted to be as creative as possible. After completing this project, I hope that when I perform future data collection, I am as accurate as possible when gathering TIME-sensitive data because missing minutes do add up in the end. I also hope that when I perform future analyses, I strive for better aesthetic appeal because I know that sometimes, the audience perusing my analyses will not be data science experts. By creating more appealing data graphics, I can appease to those who like aspects that are organized and well-kept. Whether I need to learn a new aesthetic function or a new loading package, I will strive for better visualizations. while I believe two weeks is sufficient for the questions that I answered, I believe that since life is very unpredictable, I would most likely need at least two months worth of data to see if I can find any other strong correlations between seemingly distinct events. While I have realized that correlation does not equate to causation, I can utilize these correlations to examine any confounding variables that may have influenced my results. It would not necessarily be hard to collect the data, mostly because the data involved in this project is time-sensitive rather than "Does it happen or not?" As long as I put in the effort to collect the data persistently for two months, it would not be difficult unless I succumb to a long-term emergency. As someone who provides data to others, I expect to share data in an unbiased manner, with as much purpose as possible to only inform. While I may have an inherent motive in my mind in terms of persuading my audience to see my perspective, I hope that my data speaks for itself instead of using tricky tactical means in changing data that skew in my favor greatly. As someone who analyzes others' data, some ethical responsibilities that I need to carry out include confirming that the data was collected in an accurate manner (or at least as accurate as possible), while also accounting the fact that there may be some small room for error. I also need to make sure that if data collecting involves other individuals, that they were treated in a manner that aligns with the ASA ethical guidelines so that the results that I examine are affirmed and collected without any harm.