Work Distractions

SAT 231: Calendar Query

Ziji Zhou

Last updated October 9, 2021

How Much Distractions Do I Have During Work

```
# Data import and preliminary wrangling
distractions_data <- "Distractions.ics" %>%
  # Use ical package to import into R
  ical_parse_df() %>%
  # Convert to "tibble" data frame format
  as_tibble() %>%
  mutate(
    # Use lubridate 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_sec = end_datetime - start_datetime,
    duration_sec = difftime(end_datetime, start_datetime, units = "secs"),
    # Convert calendar entry to all lowercase and rename
    activity = summary)  %>%
  # Pivot longer into start and end time          Not "tidy" variable name
  pivot_longer(c("start","end"), names_to = "Start/End", values_to = "time") %>%
  mutate(
    # Get time of day
    time_of_day = hms::hms(seconds = second(time), minutes = minute(time), hours = hour(time)),
    # Get date as a categorical variable
    date = floor_date(time, unit = "day"),
    date_chr = paste(as.character(month(date)), as.character(day(date)), as.character(year(date)), sep =
    type = NA,
    class = NA,
    work_uid = NA
  )

schoolwork_data <- "Schoolwork.ics" %>%
  # Use ical package to import into R
  ical_parse_df() %>%
  # Convert to "tibble" data frame format
  as_tibble() %>%
  mutate(
    # Use lubridate 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_sec = end_datetime - start_datetime,
    duration_sec = difftime(end_datetime, start_datetime, units = "secs"),
    # Convert calendar entry to all lowercase and rename
    activity = summary,
    # Split activity column into work and class columns
    type = substring(activity, 0, regexpr(", ", activity) - 1),
    class = substring(activity, regexpr(", ", activity) + 2))  %>%
  # Pivot longer into start and end time
  pivot_longer(c("start","end"), names_to = "Start/End", values_to = "time") %>%
  mutate(
    # Get time of day
    time_of_day = hms::hms(seconds = second(time), minutes = minute(time), hours = hour(time)),
    # Get date as a categorical variable
    date = floor_date(time, unit = "day"),
    date_chr = paste(as.character(month(date)), as.character(day(date)), as.character(year(date)), sep
  )

# Attach type of work to distractions
for(i in 1:nrow(distractions_data)){

  for(x in 1:nrow(schoolwork_data)){

    if(x%%2 == 1 && i%%2 == 1){
      if(as.POSIXct(distractions_data$time[i]) %within% interval(as.POSIXct(schoolwork_data$time[x]),as
        distractions_data$class[i] = as.character(schoolwork_data$class[x])
        distractions_data$type[i] = as.character(schoolwork_data[x,"type"])
        distractions_data$work_uid[i] = as.character(schoolwork_data[x,"uid"])
        distractions_data$class[i+1] = as.character(schoolwork_data$class[x])
        distractions_data$type[i+1] = as.character(schoolwork_data[x,"type"])
        distractions_data$work_uid[i+1] = as.character(schoolwork_data[x,"uid"])
      }
    }
  }
}
```

I think I understand what’s going on here, but additional commentary would help.

There’s a tidy-friendly package **fuzzyjoin** that has a set of functions for joining data by interval that should work well for this. (I don’t expect you to have known this)

```
# Display the data
schoolwork_data
distractions_data
activities
distractions
```

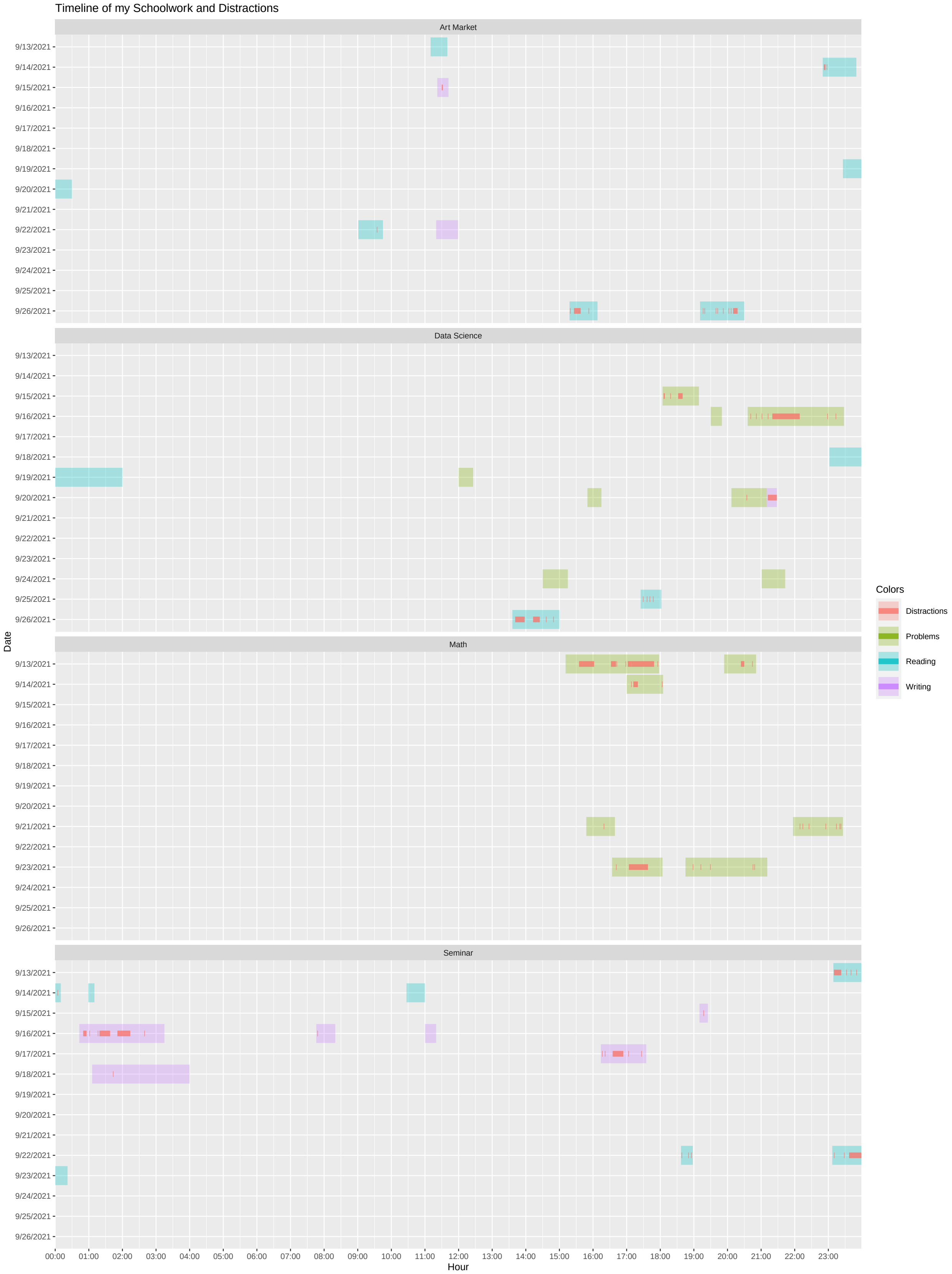
- What type of distractions do I interrupt each types of work and for which class?
- In which class am I the most productive? What type of work? What time of day? (Productiveness is defined by the amount of distractions during that period of work)

My data collection process involved logging the each time I sit down seriously to do schoolwork. I would take note the class and type of work (split between problems, reading, and writing). For each time I go away from my work I would take note of when I started going off track and when I returned as a seperate data set. My distractions were categorically split into phone, text, talk, food, web, restroom, music.

First Graph

This visualization presents my work over the two weeks of data collection. The x axis is the date and the y axis indicates the time of that day. Each type of work is divided up by color and class faceted into my four rows to help present which class am I the most productive (or as the data came out to be, unproductive). The orange streaks within the wider transparent blocks represent each distraction and the proportion of the work time it took up.

```r
# Be sure to provide meaningful title and axes labels
# Code for data visualization #1
schoolwork_data %>%
  ggplot(aes(x = as.POSIXct(time_of_day), y = reorder(date_chr,desc(date_chr)))) +
  scale_x_datetime(date_breaks = "1 hours",
                   date_labels = "%H:%M",
                   expand = c(0, 0)) +
  geom_line(aes(color = type,
                group = uid),
            size = 10,
            alpha = 0.3) +
  geom_line(data = distractions_data ,
            aes(group = uid,
                color = "Distractions"),
            size = 3,
            alpha = 0.8) +
  facet_wrap(~class, nrow = 4) +
  labs(title = "Timeline of my Schoolwork and Distractions",
       x = "Hour",
       y = "Date",
       color = "Colors")
```


Timeline of my Schoolwork and Distractions
Art Market
Data Science
Math
Seminar
Date
9/13/2021
9/14/2021
9/15/2021
9/16/2021
9/17/2021
9/18/2021
9/19/2021
9/20/2021
9/21/2021
9/22/2021
9/23/2021
9/24/2021
9/25/2021
9/26/2021
Hour
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Colors
Distractions
Problems
Reading
Writing

Second Graph

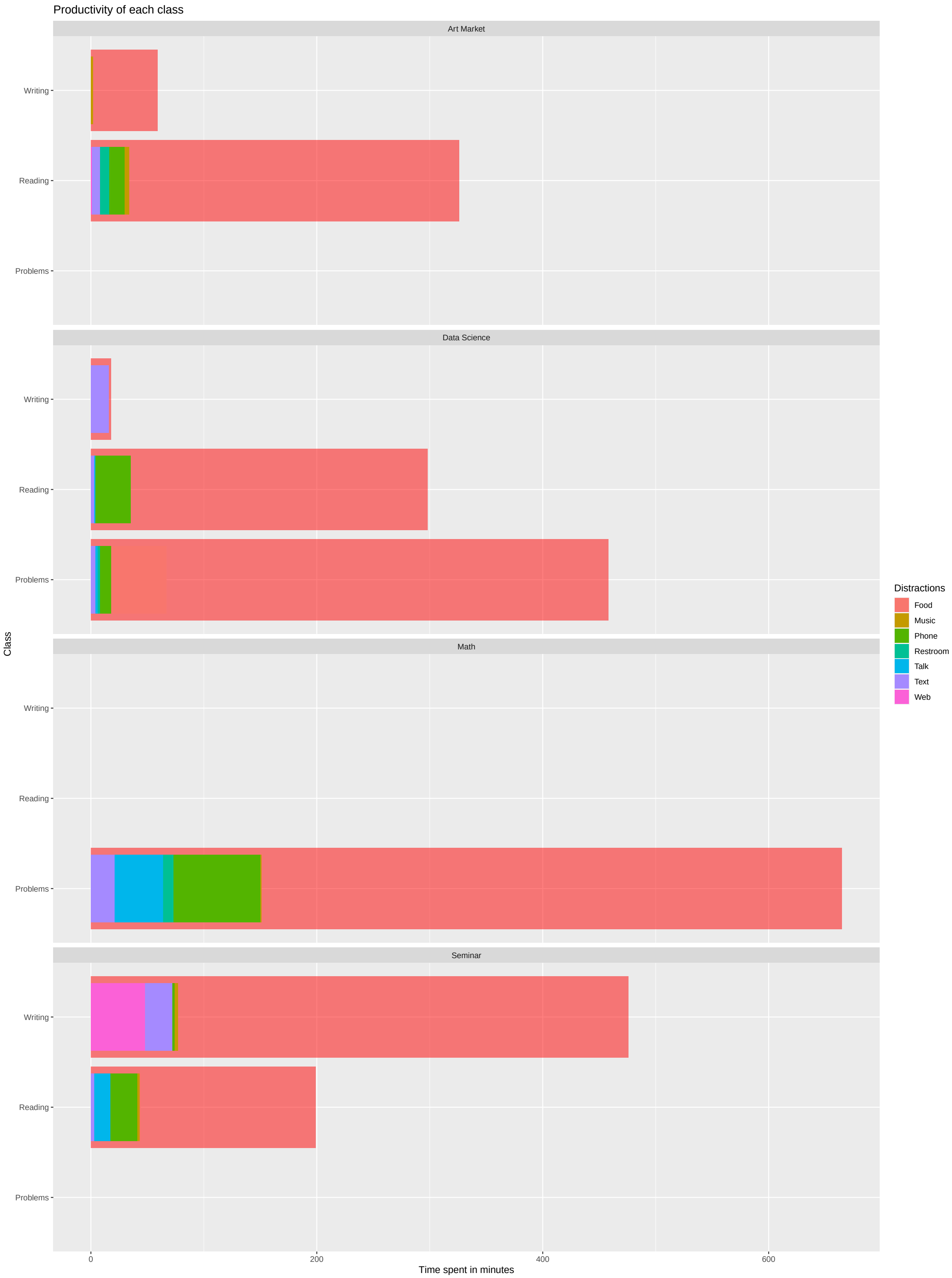
This faceted bar graph shows the distribution of my time with each class and my distractions during them. The y axis displays the minutes while the x is the class the bar represents. The graph is faceted for each type of work and the corresponding distractions. Each distraction is then divided up into color to show which type of distraction is most common for each type of work.

```r
# Be sure to provide meaningful title and axes labels

schoolwork_summary <- schoolwork_data %>%
  group_by(type, class) %>%
  summarize(duration_sec = sum(duration_sec)/2) %>%
  arrange(desc(duration_sec))

distractions_summary <- distractions_data %>%
  group_by(type, class, summary) %>%
  summarize(duration_sec = sum(duration_sec)/2) %>%
  arrange(desc(duration_sec))

# Code for data visualization #2
schoolwork_summary %>%
  ggplot(aes(x = type, y = duration_sec/60)) +
  geom_col(width = 0.9, alpha = 0.5, fill = "red") +
  geom_col(data = distractions_summary, aes(fill = summary), width = 0.75) +
  scale_fill_brewer(palette = "Set2") +
  coord_flip() +
  facet_wrap(~class, nrow = 4) +
  scale_fill_discrete(name = "Distractions") +
  labs(title = "Productivity of each class",
       x = "Class",
       y = "Time spent in minutes",
       color = "Distractions")
```


Productivity of each class
Art Market
Data Science
Math
Seminar
Writing
Reading
Problems
Class
0
200
400
600
Time spent in minutes
Distractions
Food
Music
Phone
Restroom
Talk
Text
Web

Table

My table is a summary of every individual work session and its corresponding distractions, seperated by class. The table shows the total sessions and distractions, average time per distraction and work session, and finally the proportion of time I spent each session distracted.

```r
# Code for table
schoolwork_total <- schoolwork_data %>%
  group_by(class) %>%
  summarise(total_time_spent = sum(duration_sec),
            average_work = mean(duration_sec/2))

distractions_total <- distractions_data %>%
  group_by(class) %>%
  summarise(total_time_distracted = sum(duration_sec),
            average_distraction = mean(duration_sec/2))

total_summary <- schoolwork_total%>%
  inner_join(distractions_total, by = c("class" = "class")) %>%
  mutate(proportion_distracted = as.double(total_time_distracted)/as.double(total_time_spent)) %>%
  arrange(desc(total_time_spent))

total_summary %>%
  kable(col.names = c("Class", "Total time spent", "Avg session", "Total time distracted", "Avg distrac
```

Class	Total time spent	Avg session	Total time distracted	Avg distraction	Proportion
Data Science	92880 secs	1786.15 secs	14160 secs	177.00 secs	0.15
Seminar	81000 secs	1557.69 secs	14520 secs	145.20 secs	0.18
Math	79800 secs	2850.00 secs	18120 secs	167.78 secs	0.23
Art Market	46200 secs	1283.33 secs	4320 secs	72.00 secs	0.09

Through my data collection and visualization I was able to get an idea of what leads to the most distractions. Surprisingly my reading for art market had the least amount of distractions, an explanation could be because it has the least work and every time I start reading I have an end in sight. It seems that the afternoon period is the least productive for me with lots of distractions during that period. In addition, using my phone is the most common and time consuming distraction, showing up no matter time period nor class. I think by silencing my phone and placing it far away could do wonders for my productivity. Lastly, short periods of study contains the least time spent being distracted, perhaps short burst sessions of work is better for me?

Reflections on Data Collection

Upon embarking this project I was excited at the possibility of finding patterns in my work study habits. As I was gathering ideas for what type of data to collect and what questions I can answer I realized just how much data is in our lives. The biggest difficulty was selecting the specific data points that will answer my questions before doing any actual data science work. In my case I collected the data points:

- time spent work
- which class
- what type of work (problems, writing, reading)
- time spent distracted
- which type of distraction

The process of collecting data was difficult in a few ways. Mainly I had a problem remembering to log down every time I get distracted. Considering I would be in the middle of hard work looking for some relief, it was hard initially to suddenly rememeber to mark down the exact time I had started getting off track. On top of that, the act of gathering data itself skewed my data. By going on my computer to log the time it would provide a temptation to go on other programs on my phone or computer.

I think a good way to avoid these problems in the future is to observe and collect data as a thrid party observer. By not interacting with the subject it would become a much more focused job (and thus minimizing instances where I would forget to log data) and impacting the data itself. Having a more defined question and draft of the data visualization and wrangling before even starting would help massively in understanding which data points are important to consider.

For data on economics and sports it seems sometimes that there can never be too much data. I maintain this stance as long as the collector understands which data points are more significant. I think once the data wrangling and visualization starts there can be a unforseen need to supplement the study with additional data. Sports and economics are luckily two sectors that are extremely data reliant, each with years of professionals refining the craft of data collection and a vast variety of online resources that are easy to access. The difficulty then becomes the interpretation and careful selection of the significant parts of the massive databases that are available.

When I provide data to the tech companies that exist today, I have zero expectations for privacy. The use for data is simply too lucrative and I believe it is somewhat naïve to expect anything other than almost public access to the data that I provide for the internet. Everything digital is becoming fully connected, if not already. It is then our job to utilize the data in an ethical and positive way rather than fight this new era of openness and lack of privacy.

While analyzing other's data, it is important to consider the implications of the publications. Simple mistakes like forgetting to label logartihmic scales can have far reaching effects of misinformation once shown to specific groups of people. Data is often used as a tool for agendas and we must provide enough context within our analysis to prevent the misuse of our work. Making sure that we are gathering the data ethically is also important, especially when it comes to web scraping.