Kevin Ma Calendar Schedule

STAT231: Google Calendar Analysis

Kevin Ma

Due 3/19/21

Contents

0.1 Importing The Calendar

```
#importing calendar
path <- "/Users/kevinma/Github/Stat231/Homework"
filename <- "kma8222178@gmail.com.ics"

my_calendar0 <- ical_parse_df(file = paste0(path,"/",filename)) %>%
  mutate(start_datetime = with_tz(start, tzone = "America/New_York")
    , end_datetime = with_tz(end, tzone = "America/New_York")
    , length_seconds = end_datetime - start_datetime
    , date = floor_date(start_datetime, unit = "day"))
```

```
#Data wrangling:
#Create subgroups for each type of activity, filter for relevant dates, get weekday for each date, and
my_calendar1 <- my_calendar0 %>%
  filter(date > "2021-02-24") %>%
  mutate(
    day = weekdays(date),
    sub_group = case_when(summary == "Computer Science" |
                            summary == "Thesis" |
                            summary == "Data Science" ~ "Classes",
                          summary == "Tennis" |
                            summary == "Workout" ~ "Exercise",
                          summary == "Breakfast" |
                            summary == "Lunch" |
                            summary == "Dinner" ~ "Meals",
                          summary == "Sleep" ~ "Sleep"),
    length_hour = as.numeric(round(length_seconds / 3600, digits = 2))
  )
```

The questions I wanted to answer in this assignment are the following:

1. What is the specific breakdown between the activities I do throughout the day, including classes, tennis, working out, and different meals?
2. Are there any trends during the recording period for the amount of time I spend on different types of activities?
3. For each day I recorded, how productive am I? How much time do I spend on classes versus other activities?

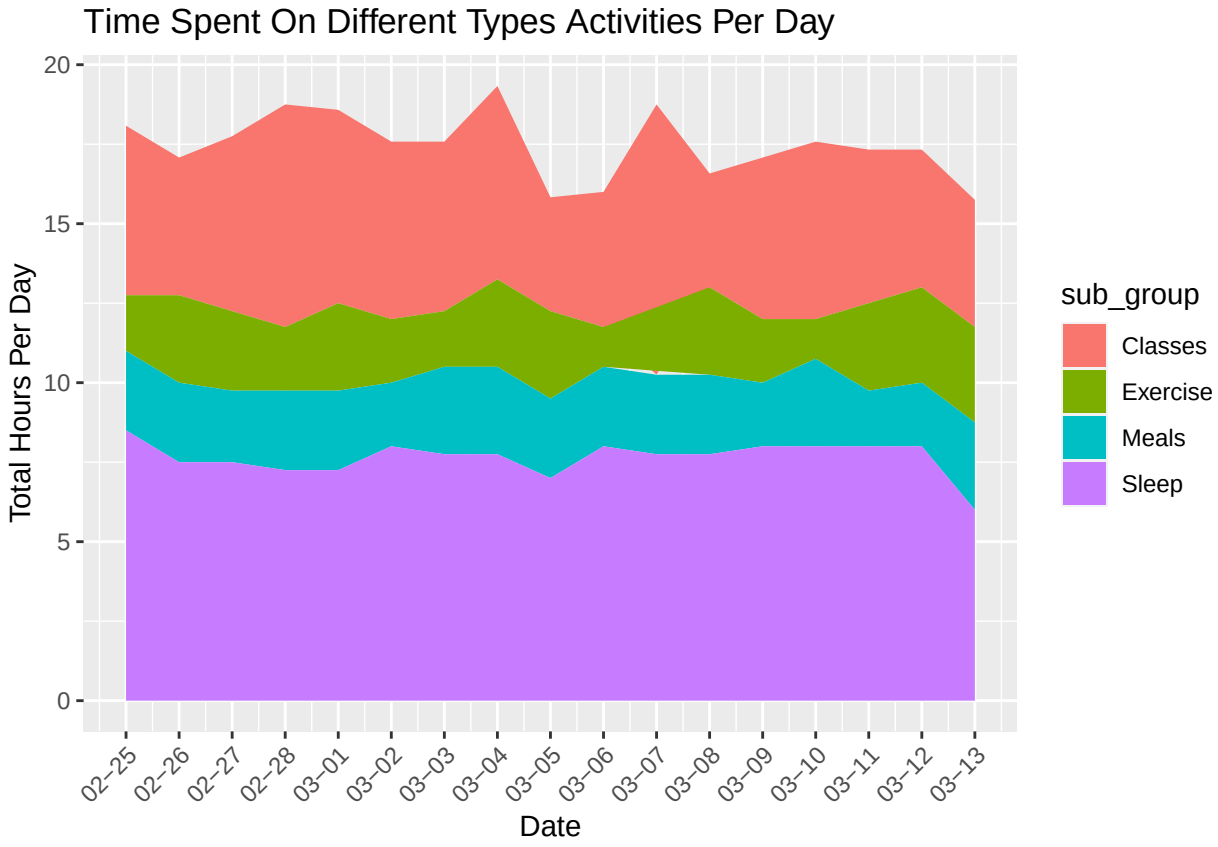
0.2 Graphic 1: Stacked Area Visualization

For this visualization, I want to present the amount of time I allocated to class, exercise, meals, and sleep for each day. I had to condense related activities into subgroups. I also had to make sure that there was data representing the total number of hours for each subgroup of activities, as some days I would play tennis and workout or had two classes. I decided to use a stacked area chart that shows the total number of hours spent on each of the four subgroup activities over the recording period.

```
#Data wrangling: Want the sub_group, the date, and the total time spent per sub_group
total_time <- my_calendar1 %>%
  group_by(sub_group, date) %>%
  summarise(total_time = sum(length_hour))
```

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

```
#Plotting the line chart
 ggplot(total_time, aes(x = as.Date(date), y = total_time, fill = sub_group)) +
  geom_area() +
  labs(x = "Date", y = "Total Hours Per Day", color = "Activity") +
  ggtitle("Time Spent On Different Types Activities Per Day") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_x_date(breaks= "days", date_labels = "%m-%d")
```



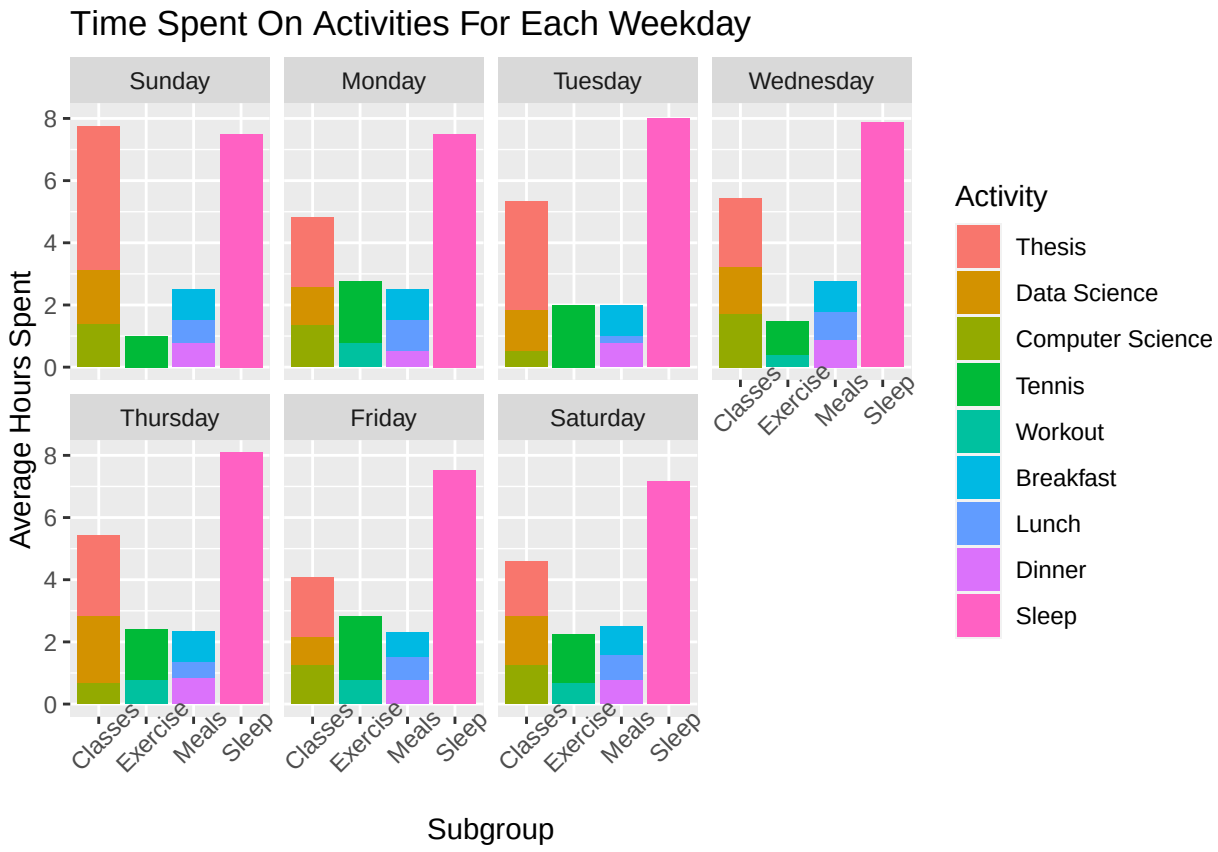
0.3 Graphic 2: Bar Graph Class Breakdown

In this visualization, I want to compare how I spend my day across different days of the week. I want to examine the total amount of time I spend on each activity, sorted by subgroup, for each day. Each panel represents a day of the week, while each stacked bar represents a subgroup. Within each stacked bar, the colors denote the activity. This format allows me to see how long on average I spend on an activity for a given day. For example, I can see that each Sunday I spend on average 4 hours working on my thesis.

```
my_calendar2 <- my_calendar1
#Order the days of the calendar
my_calendar2$day <- factor(my_calendar2$day,
                           levels = c("Sunday", "Monday", "Tuesday", "Wednesday",
                                      "Thursday", "Friday", "Saturday"))
#Order the activities in the legend
my_calendar2$summary <- factor(my_calendar2$summary,
                               levels = c("Thesis", "Data Science", "Computer Science",
                                          "Tennis", "Workout", "Breakfast", "Lunch",
                                          "Dinner","Sleep"))
#Wrangle data set: Find the average time spent on each activity. I record Sunday to Wednesday twice and
#I have to then group by sub_group, summary, and day then create the variable total time, which gives m
my_calendar2 <- my_calendar2 %>%
  mutate(
    average_lhour = case_when(day == "Sunday" |
                                day == "Monday" |
                                day == "Tuesday" |
                                day == "Wednesday" ~ length_hour/2,
                              day == "Thursday" |
                                day == "Friday" |
                                day == "Saturday" ~ length_hour/3)
    ) %>%
  group_by(sub_group, summary, day) %>%
  summarize(tot_time = sum(average_lhour))
```

```
## `summarise()` has grouped output by 'sub_group', 'summary'. You can override using the `.groups` argu
```

```
#plot the visualization
ggplot(my_calendar2, aes(x = sub_group, y = tot_time))+
  geom_col(aes(fill = summary)) +
  facet_wrap(~day, nrow = 2, ncol = 4)+
  labs(x = "Subgroup", y = "Average Hours Spent", fill = "Activity") +
  ggtitle("Time Spent On Activities For Each Weekday") +
  theme(axis.text.x = element_text(angle = 45))
```


Time Spent On Activities For Each Weekday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Average Hours Spent
0
2
4
6
8
Classes
Exercise
Meals
Sleep
Subgroup
Activity
Thesis
Data Science
Computer Science
Tennis
Workout
Breakfast
Lunch
Dinner
Sleep

0.4 Graph 3: Table

In this table, I want to summarize the amount of time spent on each subgroup for each weekday. In order to do this, I needed the total hours from each subgroup for each day, so I group by date and subgroup and use the `summarize` function to create the total_hours variable. I then use the function `pivot wider` to turn the dataset into a "wide" format. I add the "Other" variable to denote time I spend on miscellaneous activities. The ratio I create is the ratio between the time spent on classes per day versus the time spent on "other", which is a rough proxy for time spent non-essential activities. This ratio can be used as an estimate on how productive I am for each day of the recording period.

```
#Want wide data because want hours per subgroup per day
caltable <- my_calendar1 %>%
  group_by(date, sub_group) %>%
  summarise(total_hours = sum(length_hour)) %>%
  pivot_wider(id_cols = date, names_from = sub_group, values_from = total_hours)
```

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

```
#Remove NA values by changing to zero
caltable[is.na(caltable)] = 0
#Rename date column, add "other" column, add a weekday column, add hours productive ratio and classes t
new_caltable <- caltable %>%
  rename(Date = date)%>%
  mutate(
    Other = 24 - Classes - Meals - Exercise - Sleep,
    Day = weekdays(Date),
    `Classes to Other Ratio` = round(Classes/Other,
                                     digits = 2)
  )%>%
  select(Day, Classes, Meals, Exercise, Sleep, Other, `Classes to Other Ratio`)
```

```
## Adding missing grouping variables: `Date`
```

```
#Use kable package to create the table in pdf
kable(new_caltable, booktabs = TRUE, linesep = "", align = "c", caption = "Time Spent On Different Grou
  kable_styling(latex_options = "HOLD_position") %>%
  row_spec(0, bold = TRUE) %>%
  pack_rows("Week 1", 1, 7) %>%
  pack_rows("Week 2", 8, 14) %>%
  pack_rows("Week 3", 15, 17)
```

Table 1: Time Spent On Different Groups of Activities

Date	Day	Classes	Meals	Exercise	Sleep	Other	Classes to Other Ratio
Week 1							
2021-02-25	Thursday	5.33	2.50	1.75	8.50	5.92	0.90
2021-02-26	Friday	4.33	2.50	2.75	7.50	6.92	0.63
2021-02-27	Saturday	5.50	2.25	2.50	7.50	6.25	0.88
2021-02-28	Sunday	7.00	2.50	2.00	7.25	5.25	1.33
2021-03-01	Monday	6.08	2.50	2.75	7.25	5.42	1.12
2021-03-02	Tuesday	5.58	2.00	2.00	8.00	6.42	0.87
2021-03-03	Wednesday	5.33	2.75	1.75	7.75	6.42	0.83
Week 2							
2021-03-04	Thursday	6.08	2.75	2.75	7.75	4.67	1.30
2021-03-05	Friday	3.58	2.50	2.75	7.00	8.17	0.44
2021-03-06	Saturday	4.25	2.50	1.25	8.00	8.00	0.53
2021-03-07	Sunday	8.50	2.50	0.00	7.75	5.25	1.62
2021-03-08	Monday	3.58	2.50	2.75	7.75	7.42	0.48
2021-03-09	Tuesday	5.08	2.00	2.00	8.00	6.92	0.73
2021-03-10	Wednesday	5.58	2.75	1.25	8.00	6.42	0.87
Week 3							
2021-03-11	Thursday	4.83	1.75	2.75	8.00	6.67	0.72
2021-03-12	Friday	4.33	2.00	3.00	8.00	6.67	0.65
2021-03-13	Saturday	4.00	2.75	3.00	6.00	8.25	0.48

0.5 Summary of Visualizations

In the first graphic, I used a stacked area graph to visualize the trends of the amount of time I spend on different subgroups of activities over the course of the recording period. I find it particularly interesting that the amount of time I spend on classes is quite volatile, while the time I spend on exercise and meals remains fairly constant throughout the recording period. The amount of time I spent on classwork took a dive for the days of March 5th and March 6th, which were a Friday and Saturday respectively. Overall, it appears that my activity level is fairly consistent, with the dip on the last day recorded of March 13th likely mirroring the level of activity during the previous Saturday.

In the second graphic, I created a stacked bar chart to describe the average amount of time I spend on each type of activity for every weekday. What stands out to me is the amount of time I spend on my thesis relative to my other classes on Sundays. In addition, it appears that I get less sleep on Saturdays, while I do the least amount of schoolwork on Fridays and Saturdays. I think it is also interesting that the amount of time I sleep other than Saturday is fairly constant, which demonstrates that my attempt to fix my sleep schedule from last fall has been relatively successful.

For my last visualization, I made a table to show the amount of time that I spend on classes versus the amount of time I spend on other types activities, including activities that do not fall into any particular subgroup. The variable "Other" encapsulates activities such as zoom meetings with friends from home, browsing social media, playing games, as well as a variety of other activities. From the table I can see that my most productive day is Sunday while my most unproductive days are Fridays and Saturdays. I am a little surprised that the ratio between classes and "other" is below one for most of the days because during the week it definitely does not feel that way. Overall my class to other ratio fluctuates by a significant amount, and reflects when assignments are due. Much of my schoolwork is due on Monday and Friday and my most productive days are Sunday, Wednesday, and Thursday.

0.6 Reflection

I encountered numerous difficulties during the data collection and analysis process. Some of the difficulties were self-inflicted. For example, I would record my activities during the end of the day before going to bed which meant that I sometimes had trouble remembering the exact time and amount of time I spent on a certain activity. In addition, there were several spelling mistakes when I recorded activities, which made data wrangling more complicated. While I could easily fix such mistakes manually for this assignment, fixing such errors in a dataset with thousands of entries would be much more difficult. When I was analyzing data, I found that turning the dataset into the format that I wanted to be particularly challenging. While I found creating the visualizations to be mostly straightforward, manipulating the data so that it could be graphed was quite difficult.

The main hurdles I encountered when gathering data were remembering events accurately during the day, as mentioned previously, and having unspecific categories for my activities. The start and end times recorded for each activity entry were not exact, except for pre-planned activities such as class zoom meetings and tennis practice. Having inexact data would negatively impact future analysis projects because accurate data is needed for models and visualizations to be defensible to critique. Another issue I created for myself when gathering data was having vague breakdowns for my activities. For example, I didn't separate class zoom meetings from time spent on homework. In addition, I could have separated my thesis work into advisor meetings and individual work, or separate tennis practices between team and individual practices. For future projects, these inaccuracies and unspecific data entries would create difficulties.

Having such problems for future projects would both limit the types of analysis I can do and create inaccuracies. While my questions for this project were quite general, I would not be able to answer more specific questions with the type of data that I collected. In addition, having data entries that are not exact may lead to bad conclusions about my activities. For example, the ratio "Classes to Other" for February 25th was slightly below 1, however, it could very well be the case that I missed time on various activities that could increase that ratio to be over 1. Thus being more diligent when recording data and being more specific would solve most of these issues for future projects.

To answer my questions of interest, I believe that having a larger dataset and having more specific data would be helpful. Two weeks is too short to establish trends in my daily activites, and I may just be capturing noise. Having a larger sample size of around 6 weeks would provide enough data to strongly support any conclusions made. Collecting this data should not be significantly more difficult than collecting data was for this project. Being more disciplined in recording data earlier in the day and being more specific are the major challenges, but the process should be manageable.

When I provide data to companies, I have expectations that my data will remain anonymous unless I have given consent. With any data collection, I expect to have some degree of control over who owns my data and as well as who has access to it. I believe that companies should be required to notify me when they sell my data to third parties, whether it is for more personalized advertising or for government research. Unfortunately, many companies appear to recognize that people would be reluctant to allow strangers to access their data and turn to more secretive means of gathering data from their customers/users.

Similarly, if I analyze another person's data I have the ethical responsibility to consider that person's privacy. I must take into account whether the information is sensitive and can identify an individual, which would limit what I can publish and the types of analysis I can ethically conduct. For example, if I were to analyze the calendars for all of the other students in STAT-231, I have to make sure that I have everyone's consent before accessing their personal information and allow people who are uncomfortable sharing their data the opportunity to not participate. Lastly, I should always err on the side of caution and notify the relevant individuals if I plan on doing more with their data than they have previously agreed to.