YOUR TITLE HERE

STAT 231: Calendar Query

Sebastian Montesinos

Last updated March 8, 2022

Introduction

The first question I addressed is how the amount of time I spend on each class changes over the course of a week. One of my goals this semester is to have a structured and planned week where I dedicate certain days to certain classes. This allows me to easily plan when I will do work and not get overloaded on any particular day. I aim to concentrate my data science work towards the end of the week - specifically, around Thursday, Friday, and Saturday. I aim to do my oceanography and religion work around the start of the week, particularly Monday and Tuesday respectively. Finally, I aim to scatter my thesis work throughout the week.

The second question I address is how the amount of time I spend on work in a day relates to the amount of personal time I take during a day. One might expect that since the more time I spend on one activity the less time there is for the other, there will be a simple negative correlation. However, I have noticed that I sometimes give myself more time for recreational activities when I have done a lot of work as a 'reward', and spend more time on independent academic projects or looking for jobs when I do not spend a lot of time on work in a day since I feel guilty if I am not productive in a day. Therefore, it could also be the case that as I spent more time on schoolwork I also take more personal time, making the two positively correlated. I was also interested in how this relationship might change on weekdays vs weekends, since I try to give myself more of a break during the weekends.

The final question I addressed was how much time I spend doing work on weekdays versus on weekends. One of my aims at Amherst has been to have concentrated, productive weekdays so that I mostly have my weekends off. However, since I am writing a thesis this semester that has been harder to accomplish. I was interested in how productive I managed to be during the weekdays, and whether the time I spent on work during the weeks differed from on the weekends.

Data collection

I coded my data into three broad categories: 'work', for anything related to school (ie. class and homework), 'personal' for anything related to recreation (ie. gaming, seeing friends) and 'extracurricular' for any club activities. I decided to code the work category as four subcategories corresponding to my four classes: data science, oceanography, religion, and my thesis. I decided not to further divide the other two categories since my questions did not require doing so. So, in the end, I had six ways I marked off time on my calendar: 4 for my classes, 1 for personal time, and 1 for extracurricular. The units for all of these categories were in time, specifically minutes spent on each activity, which I used google calendar to code in. I did not code every single thing I did in the day, so tasks such as job applications or independent projects did not show up on my calendar.

```r
# Data import and preliminary wrangling
calendar_data <- "SMontesinosCalendarQuery.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"),
    ## Parsing dates and times
    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) %>%
  ## Turning the date variable into a date type
  mutate(date = ymd(date)) %>%
  ## Including only dates after I started collecting for the project
  filter(year >= 2022 & month >= 2 & day >= 17 | year >= 2022 & month >= 3) %>%
  ##Removing trailing whitespace
  mutate(activity = str_trim(activity)) %>%
  ##Replacing spaces with underscores in activity names
  mutate(activity = str_replace(activity, " ", "_")) %>%
  ##Creating another column that records whether the observation was made on the weekday or weekend
  mutate(week_status = case_when(day_of_week == "Sat" | day_of_week == "Sun" ~ "Weekend",
                                 TRUE ~ "Weekday"))
```

To address my first question, I intend to create a line graph that shows the average amount of time I spend on work on each day of the week, by the specific class I am working on. This graph will have the day of the week on the x axis and the time I spend on work on the y axis. There will be four lines, each corresponding to my four courses.

To address my second question, I intend to create a scatterplot that shows the relationship between the amount of work I do in a day and the amount of personal time I take in a day. The x axis will represent work time and the y axis will represent personal time. A line of best fit will be used if appropriate, and I will use faceting to create a version of the graph for weekdays & for weekends.

```r
# Preparing dataset for first visualization
# Computing total duration for each activity per date
activities_total <- calendar_data %>%
```

```
  group_by(date, day_of_week, activity) %>%
  summarize(duration = sum(duration)) %>%
#Pivoting wider to get rows as days and activities as columns
  pivot_wider(names_from = activity, values_from = duration)

#Filling in columns with NAs with '0' for 0 time spent
activities_total[is.na(activities_total)] = 0

#Calculating the mean time spent for each activity by day of the week
activities_average <- activities_total %>%
  group_by(day_of_week) %>%
  summarise(DS_Average = mean(data_science),
            O_Average = mean(oceanography),
            R_Average = mean(religion),
            T_Average = mean(thesis)) %>%
#Pivoting back longer to put the mean time for each activity in one column
  pivot_longer(-day_of_week,
               names_to = "Activity",
               values_to = "Duration")
```

```
# Preparing dataset for second visualization
activities_comparison <- calendar_data %>%
#Selecting relevant variables
  select(day, activity, duration, week_status) %>%
# Adding a unique row identifier so I can pivot
  mutate(row = row_number()) %>%
# Pivoting wider to get each activity by day
  pivot_wider(names_from = activity, values_from = duration) %>%
# Dropping row identifier
  select(-row)
#replacing NA values with Os
activities_comparison[is.na(activities_comparison)] = 0

# Grouping by day and week status
activities_comparison2 <- activities_comparison %>%
  group_by(day,
           week_status) %>%
# Calculating total school work time, personal time, and extracurricular time for each day
  summarise(work_time = sum(thesis,
                            oceanography,
                            religion,
                            data_science),
            personal_time = sum(personal),
                            extracurricular_time = sum(extracurriculars))
#
```

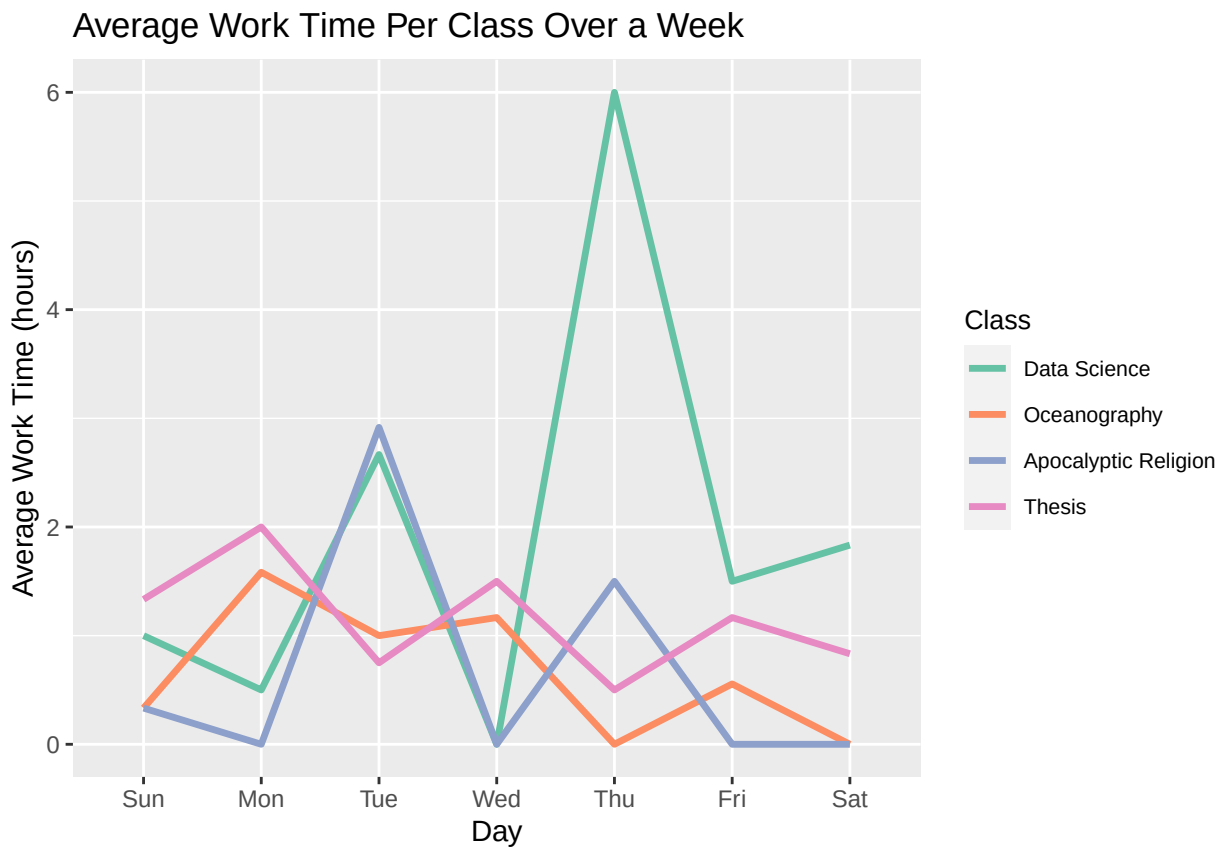
```
# Preparing dataset for table

#Grouping adjusted dataset from Vis2-prep by only week_status
week_comparison <- activities_comparison2 %>%
  group_by(week_status) %>%
#Calculating the mean, median, and standard deviation of work, personal, and extracurricular time by we
  summarise("Mean Work Time" = mean(work_time),
```

```
            "Median Work Time" = median(work_time),
            "Standard Deviation (Work Time)" = sd(work_time),
            "Mean Personal Time" = mean(personal_time),
            "Median Personal Time" = median(personal_time),
            "Standard Deviation (Personal Time)" = sd(personal_time),
            "Mean Extracurricular Time" = mean(extracurricular_time),
            "Median Extracurricular Time" = median(extracurricular_time),
            "Standard Deviation (extracurricular time)" = sd(extracurricular_time)) %>%
#Flipping the columns to rows and rows to columns to make the table nicer
  t() %>%
#Shifting the first row to be the column names
  janitor::row_to_names(1) %>%
#t() turned the data into a matrix, putting it back in a data frame format
  as.data.frame()
```

Results

My first visualization shows the average amount of time I spend on each class per day of the week. The x axis represents a particular day of the week (ie. monday, sunday) and the y axis represents the average amount of time I spend in hours on an activity. A line is used for each class, so that each point on the line at a particular day represents the average amount of time I spend on that class on that day.

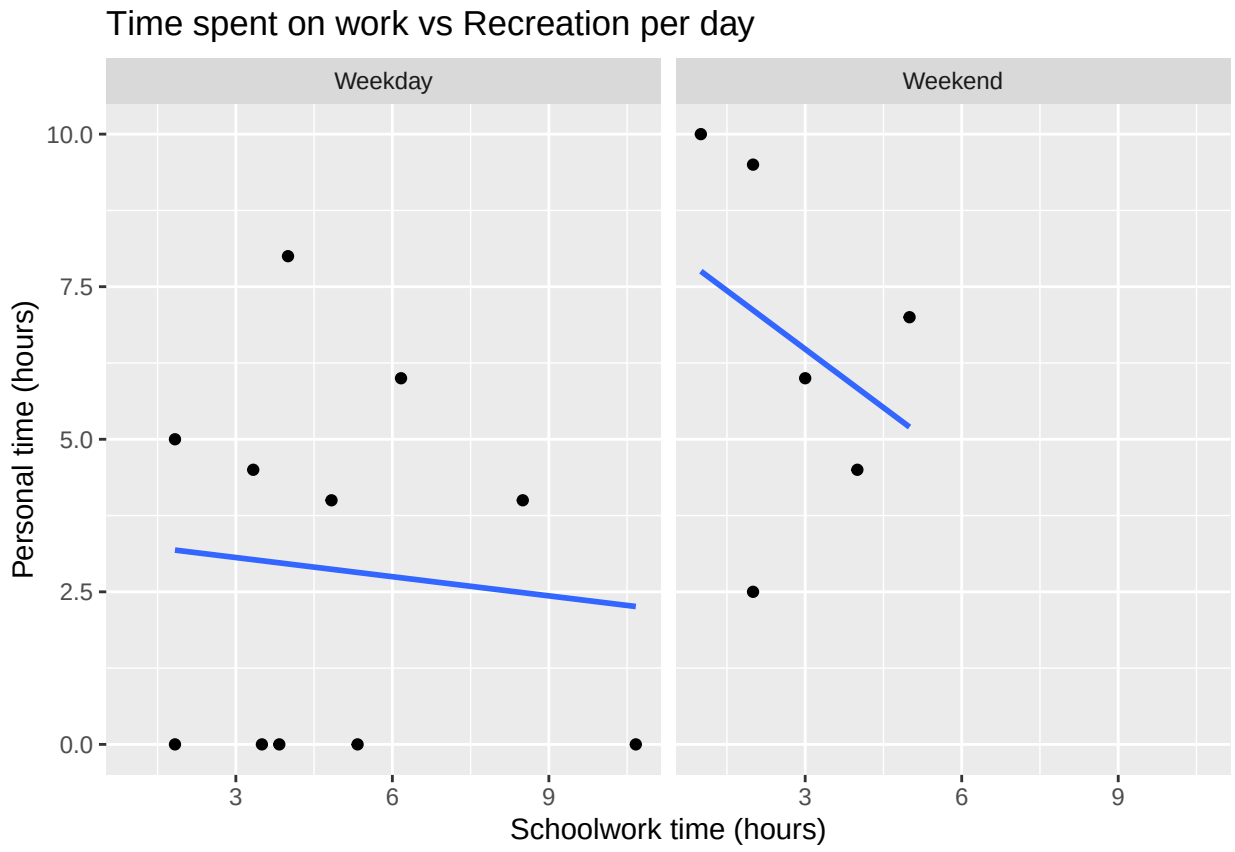
```
# Code for first data visualization
p <- ggplot(data = activities_average,
            aes(x=day_of_week,
                y = Duration,
                color = Activity,
                group = Activity)) +
  geom_line(size = 1.2, alpha = 1) +
  scale_color_brewer(type = "qual",
                     palette = 7,
                     labels = c("Data Science",
                                "Oceanography",
                                  "Apocalyptic Religion",
                                  "Thesis")) +
                     labs(DS_Average = "Data Science",
                          title = "Average Work Time Per Class Over a Week",
                          x = "Day",
                          y = "Average Work Time (hours)",
                          color = "Class") +
  theme(legend.text = element_text(size = 8)) +
  theme(legend.title = element_text(size = 10))

p
```



My second visualization shows the relationship between the amount of time I work per day to the amount of time I spend on personal time per day. The x axis represents the amount of work per day, and I was interested in how this impacted the amount of personal time I took, which I put on the y axis. I used faceting to divide this data into two separate graphs for the weekend and weekday.

It appears that there is no clear relationship between the amount of work time I take in a day and the amount of personal time I take in a day, lending support to neither of my hypotheses. Because of this lack of a clear relationship, I did not add a line of best fit to the graphs as it would not be appropriate. It does appear that there may be a slightly negative relationship between personal time and work time on the weekends, but I would need more data to draw any significant conclusion.

```
#Plotting total work time against personal time per day
n <- ggplot(data = activities_comparison2, aes(x= work_time, y = personal_time)) +
#Creating 1 plot for weekends and one for weekdays
  facet_wrap(~week_status, nrow = 1) +
#Using points to represent each day
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
#Labeling the graph
  labs(title = "Time spent on work vs Recreation per day",
       y = "Personal time (hours)",
       x = "Schoolwork time (hours)")
n
```




```
# Code for table
# Only code for your table should be here (no or very minimal wrangling code)

week_comparison <- week_comparison %>%
  kable(booktabs = TRUE, digits = 2, caption = "Differences in time Spent on Weekdays vs. Weekends") %>%
  kable_styling(latex_options = "HOLD_position")

week_comparison
```

Table 1: Differences in time Spent on Weekdays vs. Weekends

	Weekday	Weekend
Mean Work Time	4.893939	2.833333
Median Work Time	4.0	2.5
Standard Deviation (Work Time)	2.705214	1.471960
Mean Personal Time	2.863636	6.583333
Median Personal Time	4.0	6.5
Standard Deviation (Personal Time)	2.950347	2.888194
Mean Extracurricular Time	0.7272727	1.0000000
Median Extracurricular Time	0.0	0.5
Standard Deviation (extracurricular time)	1.009050	1.264911

Conclusions

I learned that I have succeeded fairly well in my goal of planning out my week such that I do work for each class on a pre-planned part of the week. I tend to do most of my data science work at the end of the week, my religion & oceanography work at the start of the week, and my thesis work scattered throughout the week and weekend. I also learned in this analysis that I spend the most time on data science rather than my thesis, which was somewhat surprising. However, I did collect my data during two non-representative weeks, something I will address in my reflection.

Next, I learned that there is no strong relationship between how much time I spend on work and how much personal time I take in a day. The idea that I give myself a break when I do a lot of work was not confirmed by the data, suggesting that I d

Reflection

Future data projects

I think one of the things that would have helped with my second analysis is coding in not only personal time but the time I spent on non-school work like independent projects or looking for jobs. Without having explicitly coded in my non-school work time, I was only able to indirectly measure that time via personal time, which was not ideal.

Did I have enough data?

I think I did not have quite enough data to completely answer some of my important questions. For instance, my first question was about how the time I spend on each class changed throughout the course of an average week. However, I only had two weeks worth of data to average across. These two weeks are not perfectly representative of my average week. For instance, I have spent a lot more time on my thesis in the past and happened to have less work for it in the last two weeks since I am between writing sections. To get a representative sample of my typical week I think collecting data for at least 1-2 months would be ideal. Collecting this data would not be too difficult since I would just need to continue to track my time for a full month.

Ethical Responsibilities

When analyzing others data, I have the responsibility to check whether I have permission to use their data and to respect the privacy of the people referenced by the data whenever possible. For instance, we should anonymize data for people who are uncomfortable with their personal information being shared in data analyses. I also have the responsibility to work to mitigate biased analyses wherever possible. This means not only avoiding analyses that use factors such as race or gender in unscrupulous ways, but also being vigilant in avoiding 'proxies' for these categories that end up producing results based on the legacy of discrimination in this country.