How is it Already 3AM?

STAT 231: Calendar Query

Bilal Tariq

2024-02-28

Introduction

We all are familiar with the same loop: you wake up ready for a brand new day at 9am with so much energy, and all of a sudden you look at the clock it's 2am and you still haven't finished your physics assignments. Where does all your time go? More specifically, how is an average college student's time (like mine) divided between each of his courses, work routine and sleep, and in which locations? These are the questions that I aim to tackle in this report.

But why is knowing where our time is spent important in the first place? I realize that over the course of a semester our lives become rather disorganized and haphazard. We also get stuck in the same routine, in the same places, over and over again. By looking at ***where*** I spend my time and ***how,*** I am able to then consider how I could optimize work efficiency and change up where I work; this would in turn allow me to have more time for friends and activities and meet new people in new places.

Data collection

For the duration of the two weeks, I made sure to record the duration of my "work" and sleep at the end of each day. In particular, my work was subdivided into 3 categorical variables: courses, homework, and my job. The variables for courses and homework were further subdivided into 4 categorical variables i.e which course was the work being done for. (For instance, Physics124 Coursework and Stat231 Homework). These were meant to record two things:

1. Location Data (where did this event take place?)
2. Interval Duration (how long did this event last?)

```
# Data import (requires **ical** package)
cal_import <- ical_parse_df("stat231tracking.ics")

# Data wrangling
mycal <- cal_import %>%
  # Google Calendar event names 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)
    across(c(start, end),
           .fns = with_tz,
           tzone = "America/New_York"),
    # Compute duration of each activity in hours
    duration_hours = interval(start, end) / hours(1),
    # Examples of getting components of dates/times
    # Note:
    # i. these could be based on either start datetime or end datetime
    # ii. you do NOT need all of these!! so only use what you need
    date = date(start),
    year = year(start),
    month_number = month(start),
    month_label = month(start,
                        label = TRUE,
                        abbr = FALSE),
    weekday_number = wday(start),
    weekday_label = wday(start,
                         label = TRUE,
                         abbr = FALSE),
    hour = hour(start),
    time = hour(start) + minute(start)/60,
    # Convert text to lowercase and remove repeated or leading/trailing
    # spaces to help clean up inconsistent formatting.
    across(c(activity, description),
           .fns = str_to_lower),
    across(c(activity, description),
           .fns = str_squish)
  ) %>%
  # The first Google Calendar entry is always an empty 1969 event
  filter(year != 1969)
```

To analyse my data, I focused on 3 key aspects of the ICS file:

1. The time interval that has been calculated using the code above
2. The location data
3. The activity name

For my first visualization, I decided to make a box plot to encapsulate the median duration of work that was spent on assignments for each class and job. This demonstrates how my time awake was spent outside of class on my work. For the second visualization, I made a bar graph showing the cumulative time spent on each of my courses, my job and sleep. As for the table,

```
# I have divided my wrangling code into various wrangling sections. The first
# one below is general wrangling code that allows me to select the columns I
# need for my first visualisation and the second one
# (i.e duration spent on activities at home & cumulative time on activities)

wrangled_mycal <- mycal %>%
  select(activity, description, duration_hours) %>%
  rename(
    Activity = activity, Location = description, Duration = duration_hours)

wrangled_mycal$Duration = round(wrangled_mycal$Duration, 3)


#wrangling for first visualisation
#We just want homework and the job records, so we search for these
#in our Activities column

homeActivities_plot <- wrangled_mycal %>%
  filter(
    Activity %in% c("job",
                    "math211homework",
                    "physics124homework",
                    "cosc112homework",
                    "stat231homework")
  ) %>%
  na.omit()

#Making sure to omit any entries where there was no coursework or
#data was incomplete or missing

#Our second visualisation is in the form of a bar chart, so we must wrangle
#the data which allows us to group the coursework and homework together

barChart <- wrangled_mycal %>%
  mutate(Activity = case_when(str_detect(Activity, "physics") ~ "Physics",
                              str_detect(Activity, "math") ~ "Math",
                              str_detect(Activity, "cosc") ~ "Comp Science",
                              str_detect(Activity, "sleep") ~ "Sleep",
                              str_detect(Activity, "job") ~ "Job",
                              str_detect(Activity, "stat") ~ "Statistics",
                              TRUE ~ "Other"))

#What we want is a barchart that merges both my coursework and homework into
#the same activity, so I used the function 'str_detect' followed by the primary
#keyword which changed the names of those entries in Activities to be the same.
#This would then allow me to group by each course when manipulating the data.

barChart_summarized <- barChart %>%
  group_by(Activity) %>%

  #Here we're grouping by activity so that we can get the
  #cumulative amount of time spent on each of the activites.

  summarize(
    Total_Time = sum(Duration) #Summation of the duration of all the activites
  )
```

```
#In the following code, I specifically am wrangling it for the table.
#Notice how I'm piping in mycal and not wrangled_mycal as I did above because
#this time round we need the 'weekday' field

wrangled_table_mycal <- mycal %>%
  select(
    activity, duration_hours, weekday_number
  ) %>%
  #Here I'm attempting to do a similar wrangling process as the bar chart,
  #where I group together the coursework and homework time and then filter
  #all rows that aren't 'sleep'
  rename(
    Activity = activity, Duration = duration_hours,
    Weekday = weekday_number) %>%
    mutate(Activity = case_when(str_detect(Activity, "physics") ~ "Physics",
                              str_detect(Activity, "math") ~ "Math",
                              str_detect(Activity, "cosc") ~ "Comp Science",
                              str_detect(Activity, "sleep") ~ "Sleep",
                              str_detect(Activity, "job") ~ "Job",
                              str_detect(Activity, "stat") ~ "Statistics",
                              TRUE ~ "Other")) %>%
    filter(Activity %in% c("Job", "Physics", "Math", "Comp Science"
                          ,"Statistics"))

summary_table_mycal <- wrangled_table_mycal %>%
  group_by(Activity, Weekday) %>%
  #Grouping by Activity and Weekday so we can focus on the
  #average duration spent on an activity per day
  summarize(
  Mean_Time = mean(Duration)
  ) %>%
  arrange(Weekday) %>%
  mutate(Weekday = case_when(str_detect(Weekday, "1") ~ "Monday",
                             str_detect(Weekday, "2") ~ "Tuesday",
                             str_detect(Weekday, "3") ~ "Wednesday",
                             str_detect(Weekday, "4") ~ "Thursday",
                             str_detect(Weekday, "5") ~ "Friday",
                             str_detect(Weekday, "6") ~ "Saturday",
                             str_detect(Weekday, "7") ~ "Sunday"))

summary_table_mycal$Mean_Time = round(summary_table_mycal$Mean_Time, 1)

#Again, using the case_when with str_detect to substitute numerical days with
#the names of the days
```

Results

So what does my day look like when I'm not in class?

Below (Fig 1.0) are box plots for each of the activities (excluding coursework and sleep) that were carried out throughout the two weeks. The activities are shown categorically on the y-axis and the duration (in hours) is shown as a quantitative variable on the x-axis. Each box plot describes 3 key things: the median duration for the activity, the interquartile range and the minimum and maximum duration of time that has been spent on the activity. From the plot, it seems as though I generally spent the most time doing my job, where a vast majority of my time was concentrated in the science center. As far as my homework, the most time was taken up by Physics which I primarily worked on in my dorm for a median of 3 hours at a time.

```
#The ggplot code alongside geom_boxplot allows us to plot a boxplot.
#aes determines what our x and y coordinates will be
#faceting allows us to divide the data by location into separate graphs
#labs allows us to label our axes

ggplot(homeActivities_plot, aes(x = Activity, y = Duration)) +
  geom_boxplot(fill = "white", color = "black") +
  theme_minimal() + #Minimalism
  coord_flip() +
  labs(x = "Activity", y = "Duration (hours)",
      title = "Distribution of Duration for Each Activity") +
      facet_grid(~Location) #faceting
```

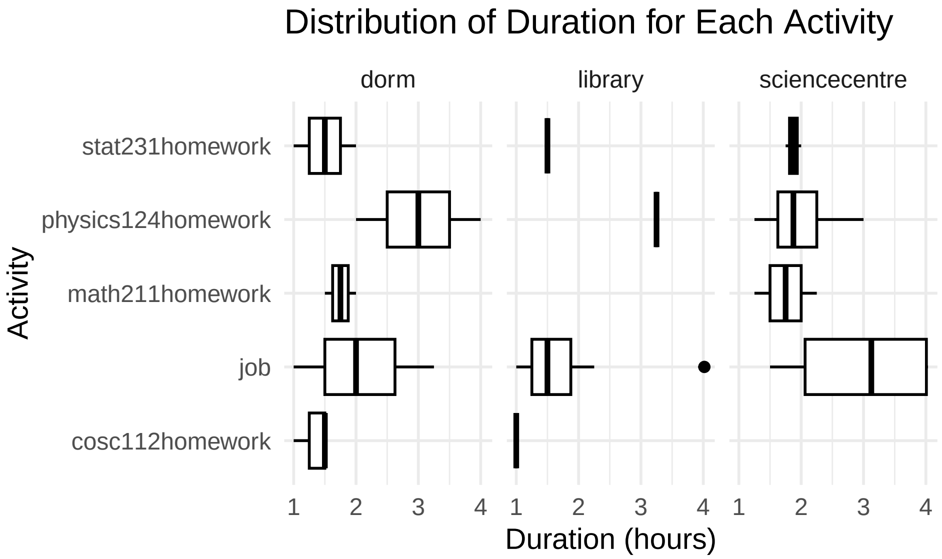

Fig 1.0

Although it was interesting to see how my activities outside of class varied in their duration, I was curious to know *exactly* what my time distribution was including all my activities over the course of the past 2 weeks. This visualization is shown in the bar chart below, where on the x - axis we have the categorical data (the activities that time was spent on) and on the y - axis we have the total duration of time spent on each activity. The visualization tells me that I was able to get over a 100 hours of sleep during the 2 weeks (shocking), and that I spent the most time on Physics (approximately 35 hours) with a close rival being my on campus job.

```
#The ggplot code alongside geom_bar allows us to plot a barchart
#aes determines what our x and y coordinates will be
#labs allows us to label our axes

ggplot(barChart_summarized, aes(x = Activity, y = Total_Time)) +
  geom_bar(stat = "identity", fill = "white") +
  labs(x = "Activity",
       y = "Total Duration (hours)",
       title = "Total Time Spent on Each Activity") +
  theme_dark() #DarkMode has been set
```

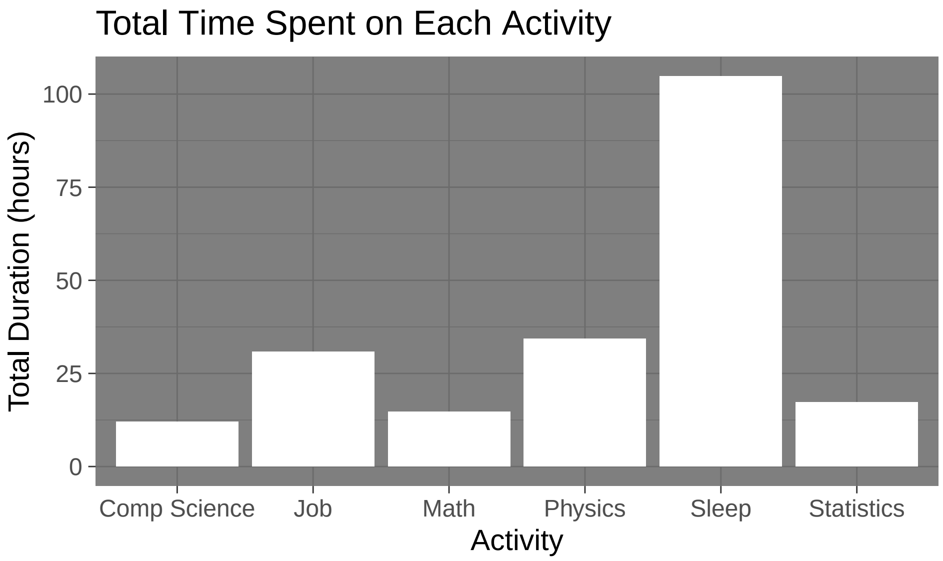

Fig 1.1

Below (table 1.0) describes the mean time I spent on my courses and my job per day of the week over the two weeks. You may notice that there are some empty entries, and that is because I did not work on those activities during those days, so they can be read as “0 hours”. I find it interesting to note how I procrastinate and leave a lot of my work for Sunday, where I spend 4 hours on average on my job and 3.6 hours on physics. You may also notice a trend with average hours spent when it comes with respect to deadlines. For statistics, I have deadlines on Sunday and Thursday nights, which is why you’ll see more time spent on those days than any other.

```
#We're using the wrangling that we did above for summary_table_mycal

pivoted_summary <- summary_table_mycal %>%
  pivot_wider(
    names_from = Weekday,
    values_from = Mean_Time
  )
```

```
#pivoting allows us to get more columns and thus divide them up into
#weekdays

kable(pivoted_summary, booktabs = TRUE)
```

Activity	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Job	1.5	2.8	1.6	2.3	1.5	2.5	4.0
Math	2.0	1.2		1.1		1.2	
Comp Science		1.0		1.1		1.0	1.5
Physics		1.8	1.9	1.0	2.3	1.0	3.6
Statistics		1.5	1.0	1.6	1.5		1.9

```
#Booktabs allows us to make our table look pretty
```

Table 1.0

Conclusions

Here is the part where we know the answer to our question "How is it Already 3am?". More specifically, we finally have the answers to which courses take up the most time and how it's split throughout the day, according to which location. I noticed my day is very scattered and disorganized (being a STEM student and all). But more importantly, I realized that I don't do much besides travel to the science centre, my dorm and library. Going forward, I'm going to make it my priority to work more efficiently rather than procrastinate till the last day and work then. Additionally, I'm going to diversify my locations of working, maybe visiting the Lyceum or Faryweather.

And the most important takeaway: I need to drop physics.

Reflection

Difficulties in Data Collection & Future Projects

Let's start off by stating the obvious: data collection is a **difficult** process. Mainly, it was the problem of making sure that my data was accurate and in line with what the goals of my questions were. My first hurdle was ensuring that I was able to track my time precisely on the calendar - this was difficult to do as often times I would find myself having gone through half the day and remembering that I hadn't tracked my day yet. Luckily, I had solutions to this: I tracked my sleep regularly on my Apple Watch which gave me an accurate representation of sleep as I would not be in a state of mind to note it down when I woke up and secondly, I had a calendar of my courses and time tracking for my job so I could determine where I spent those hours. Unfortunately, my homework time tracking was still somewhat inaccurate due to the forgetfulness of human nature and human error. My second issue with my data tracking cycle was about a couple days in I realized that I had only tracked my time in activities outside of my classes as "work". This was not helpful to me, as I wanted to know how my work outside my classes was split over the day, and so I had to revisit my data analysis cycle and add to my tracking on how much time I spent on specific homework.

If I were to repeat this project, I would focus on making sure that I had a more formalized and rigorous proposal. Although I did do it this time around, for next time I would explicitly define my variables into categorical and quantitative variables and plot dummy visualizations so I could predict the difficulties in the analysis part (i.e wrangling and coding). Sometimes, we need data in the analysis that we didn't realize was relevant in the original proposal, and so going through the process initially would help crystallize a pathway moving forward.

Was the data enough?

After I had plotted my table and visualizations, I came to a stark realization: my dataset wasn't expansive enough. There could have been biases in my data, as these past few weeks have been leading up to midterms it would mean I would have been studying more and sleeping less. It only captures a small segment of how my school semester is spent, and to accurately see the distribution of my time I would have to collect data over the period of the entire semester. This isn't to say my data didn't satisfactorily answer my questions, rather there are more questions that arise from the data I've already collected. I believe collecting data over the period of a semester would not be too difficult if one makes it a routine, and it is definitely plausible as it's simply time tracking.

Providing Data & Ethical Responsibilites

Data is our modern currency, and so just like a transaction with anyone, I have expectations from that entity. By giving out my data, my most important expectation would be for it to only be used for the activity that I have consented to, for example Google maps only using my location for the sole purpose of improving its maps features. If however, they sell my data to other companies so they can use it for targeted ads, it would nullify the contract and my expectations. These are the same ethical responsibilites that apply to me as well, where I can only use data with what people have consented for me to use them. Additionally, I should be able to raise questions and awareness when data actively discriminates against someone's identity like race, religion or gender.