MATH 154 - HW1 - Playing the Whole Game

Ra-Zakee Muhammad

due: Thursday, Sep 9, 2021

HW advice

General notes on homework assignments (also see syllabus for policies and suggestions):

1. please be neat and organized which will help me, the grader, and you (in the future) to follow your work.
2. write your name on your assignment.
3. please include at least the number of the problem, or a summary of the question (which will also be helpful to you in the future to prepare for exams).
4. it is strongly recommended that you look through the questions as soon as you get the assignment. This will help you to start thinking how to solve them!
5. for R problems, it is required to use R Markdown (or R Sweave)
6. in case of questions, or if you get stuck please don't hesitate to email me (though I'm much less sympathetic to such questions if I receive emails within 24 hours of the due date for the assignment).

Homework assignments will be graded out of 5 points, which are based on a combination of accuracy and effort. Below are rough guidelines for grading.

[5] All problems completed with detailed solutions provided and 75% or more of the problems are fully correct. **Additionally, there are no extraneous messages, warnings, or printed lists of numbers.**
[4] All problems completed with detailed solutions and 50-75% correct; OR close to all problems completed and 75%-100% correct. **Or all problems are completed and there are extraneous messages, warnings, or printed lists of numbers.**
[3] Close to all problems completed with less than 75% correct.
[2] More than half but fewer than all problems completed and $> 75\%$ correct.
[1] More than half but fewer than all problems completed and $< 75\%$ correct; OR less than half of problems completed.
[0] No work submitted, OR half or less than half of the problems submitted and without any detail/work shown to explain the solutions. You will get a zero if your file is not compiled and submitted on GitHub.

Motivation

A typical refrain from data scientists in the wild is that 80-90% of their time is spent wrangling data.

But for data scientists in the classroom, it's hard to create in-the-wild settings which require working with the entire data science process. In HW1, the student is required to create the data, wrangle the data, and visualize the data. The data you will be asked to collect and visualize are based on a Google calendar.

Assignment

1. As a preliminary step, listen to the Compromised Shoe Situation episode of Not So Standard Deviations, http://nssdeviations.com/71-compromised-shoe-situation. Note: the conversation is about 45 minutes long. Listen while folding laundry or walking to class! [Roger Peng blogged about it, but the thought process in the actual podcast is better for understanding the assignment at hand. Blog here: https://simplystatistics.org/2019/01/09/how-data-scientists-think-a-mini-case-study/.]

TODO: listen. nothing to include right here.

2. Ask yourself a question which might be answered by a calendar. How much do I study per week? or How much am I sleeping? or How much am I exercising? (Feel free to just make up the question and the data if you aren't comfortable putting your calendar into your HW assignment ... but the HW repo is private, FYI.)

TODO: How much time do I spend walking per day each week?

3. Create the dataset:
    - In a Google calendar, under "Other calendars" click on the "+" to add a calendar. Name it whatever you want.
    - Add events to your calendar. They can be as personal or impersonal as you want, but the events will be summarized in your HW assignment.
    - Tag the events using "type: event" syntax (so as to be consistent with the code below). Create at least 20 events with at least 4 types.
    - Try to put as many items on your calendar as possible. The events may be activities (what you are doing), locations (where you are), or any other creative way that you want to partition your day.
    - Go to Google Calendar -> Settings -> "Import & Export" -> Export. Unzip the file and put the dataset (with the .ics extension) into the folder associated with the GitHub repo for this assignment (the folder will be named something like ma154-hw1-yourusername).

For example, the calendar which has been exported to be included with this assignment can be viewed here:
https://calendar.google.com/calendar?cid=dTZpMW1pOHRyZXJlZG1wZTUyNmIzcXY3ajRAZ3JvdXAuY2FsZW5kYXIuZ

TODO: create the dataset. nothing to include right here.

4. Upload the dataset into R. (Note that the `ical` package will read in files from calendars other than Google. Feel free to use any calendar system you want!) Note that in the `Console` (look at the window *below*!) you'll need to import the two packages we used: `install.packages("tidyverse")` and `install.packages("ical")`. Install the packages in the console, not in the Rmd file. Then keep the two `library()` functions in the R chunk as indicated.

```
library(tidyverse)
library(ical)

# Load ics file using ical package
daily_events <-
  "whatever3.ics" %>%
  ical_parse_df() %>%
  as_tibble()
```

TODO: change the name of the file you are importing. remember to include (commit & push) the data file in your GitHub folder / repository which is part of the reproducible analysis.

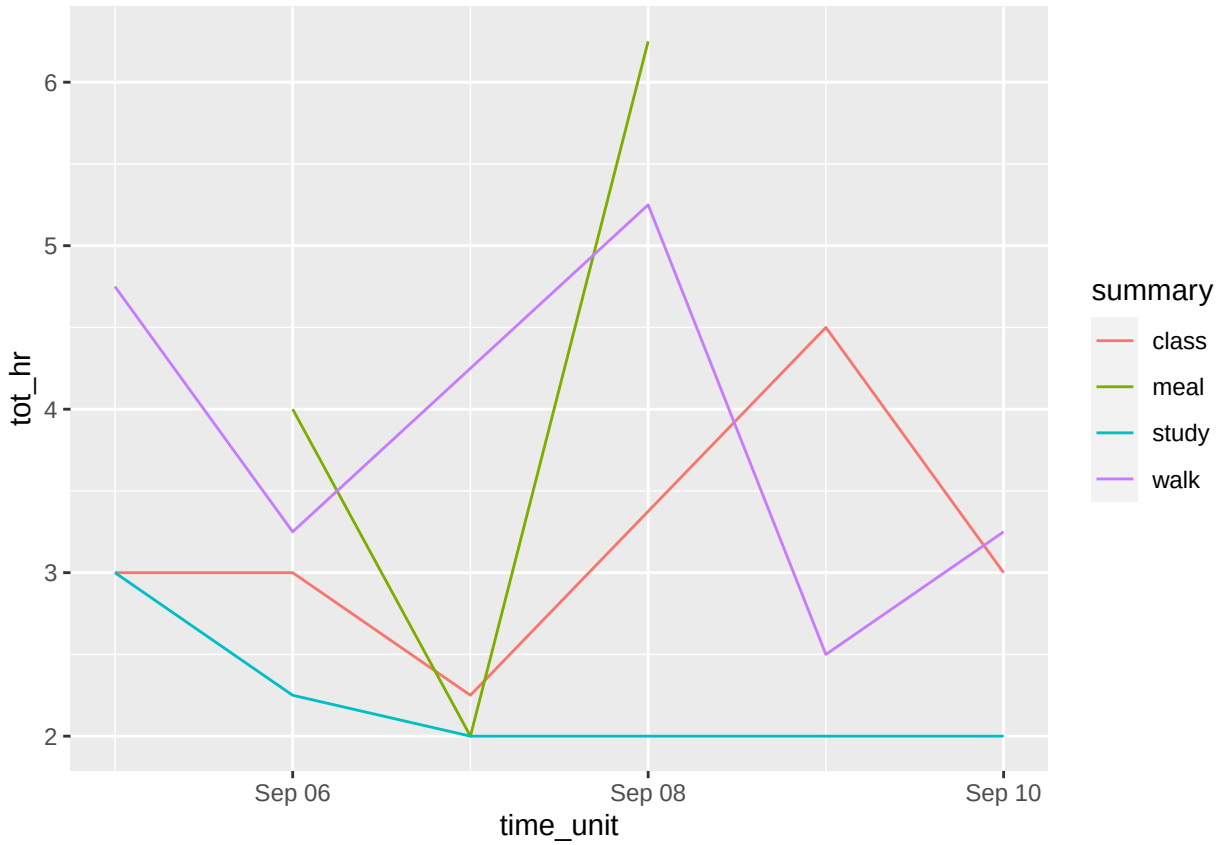
5. Wrangle, visualize, and summarize the results of the dataset. It is likely that you can just run the code below. If you want to plot a longer time period, change the date in the filter command. If you want to do something else, feel free to ask me the R code, and I will just give you the code (or ask on Discord!).

TODO: feel free to directly use the R code to wrangle & visualize (or expand it if you want to!). write a sentence or two to summarize.

```
daily_events <- daily_events %>% as_tibble() %>%
  mutate(
    start_datetime = lubridate::with_tz(start, tzone = "America/Los_Angeles"),
    end_datetime = lubridate::with_tz(end, tzone = "America/Los_Angeles"),
    hours = end_datetime - start_datetime,
    time_unit = lubridate::floor_date(start_datetime, unit = "day")
  ) %>%
  # I only want information since September
  dplyr::filter(time_unit > "2021-09-01") %>%
  # Separate out summary and subtask
  separate(summary, c("summary", "subtask"), sep = ":") %>%
  select(-subtask)
```

Wrangle

```
daily_events %>%
  # Summarize time spent
  group_by(time_unit, summary) %>%
  summarize(tot_hr = sum(hours) %>% as.numeric()) %>%
  # Plot as a line graph
  ggplot(aes(x=time_unit, y = tot_hr, color = summary)) + geom_line()
```



Visualize

Summarize This graph provides the total number of hours per date spent on a particular task. These total hour nodes are connected using line graphs. Between 3 and 2 hours are spent daily studying during the week, between 2.25 and 4.5 hours daily are spent in class, between 2 and 6.25 hours are spent daily on meals, and between 2.5 and 5.25 hours are spent daily on walking.

6. Reflect on the process by answering the following questions:

TODO: answer each of the questions below in your own words.

(a) What types of questions was Hilary trying to answer in her data collection? What types of questions might you be able to answer in your data collection?

What range of times does it usually take for Hilary to get to work, and how does commute methods influence these ranges? What time should she leave from home based on these commute times? How frequently does she use one commute method over another?

If I am measuring the duration of my walking time to classes per day of the week, I can answer the question of what days I walk the least to see if I can supplement the absence of walking with some other physical activity in order to remain in shape. Alternatively, I might use those days for workouts precisely because I wont be as tired on those days from walking to class in the sun. What days are optimal for work out given time spent walking?

(b) Name two of Hilary's main hurdles in gathering accurate data. Name two of your main hurdles in gathering accurate data.

Talking to someone during her commute and as a result her not focusing on arruving at her destination in the quickist time. The wednesday in the bay area is always without work meaning that there are fewer opportunities for hillary to meaure her commutes which is another hurdle. For me a main hurdle might be me having to stop at my dorm to pick up something (like my glasses) that I forgot to bring with me at the beggining of the day. This will add an additional amount of time onto my walking and reduce the accuracy of my measurments. On the other hand, at some point I may have to take a detour to my classes due to construction or unintentionally which may add more to my travel time.

(c) Write a few sentences contrasting data collection that is "high touch" (manual) versus "low touch" (automatic). When is high touch more accurate, when is low touch more accurate (provide an example or two)?

A particular example of the difference between high touch and low touch data is discussed during the podcast when Hilary discusses whether or not to use Google maps location history to log her travel times. Using google maps' predictive tool for commute times in this scenario would be an example of manual data collection but there is high uncertainty with this method of data collection so instead, Hilary opted for a low touch method of measuring systemnatically how long it would take for her device to transfer wifi connections from home to work. This method allowed for more certainty (not being affected by disconnections from google maps).

(d) What additional variables (covariates) did Hilary discuss as important to the data analysis but difficult to collect or not collected in her first foray of analysis? What additional variables (covariates) would you want to collect the next time you tried to understand how you spend your day?

Additional variables that Hilary discussed as difficult to collect were commute method as she had not found any automatic method for documenting whether she was taking the subway or the bus a particular day. Also, the fact that her phone had to be on for the automatic data collection of wifi transfer to go into effect was something she had not taken into consideration in her data analysis and this added additional time to some data points. In my case, for travel data in particular, I would like to take into consideration the temperature and/or weather as that most likly might affect how much time I spend walking a particular day. For instance if its raining, I might spend more time travelling than usual and I'd want to take this into consideration. Additionally, I might take into consideration whether or not I am entering a space at peak hours which might influence how quickly I can move through an area past other people.

(e) How much data do you think you'd need to collect in order to answer your research question of interest? Would it be hard to collect that data? Why or why not?

I think I would need at least 5 weeks of data so that I can compare walking times by the days of the week (at least 5 measurements per day of the week) and calculate variance by day. I don't think it would be too difficult collecting data though I believe I would probably use a manual method of logging data. I would simply have to note the time I leave from one place and the time I arrive at another throughout the day, and besides remembering to do so I wouldn't think of this as too difficult to keep up with.