Project 1: Calendar Query

Due by midnight Wednesday, February 28th

Calendar Query - Where does my time go?

This project will let you explore how you are spending your time, along with thinking about limits on personal privacy, while practicing data wrangling and the data analysis cycle.

In this **individual project**, you will undertake the full data analysis cycle to examine aspects of how you spend your time based on an analysis of data from your personal calendar. You may learn useful information that could affect how you spend your time later in the semester. Oftentimes, our expectations don't match reality, but you won't notice that unless you actually study it.

Project details

1. Pose question(s) of interest

In order to practice the data analysis cycle, you need to have question(s) of interest that you'll collect data to answer. Everyone will track their time for at least 14 days, so bear that in mind.

Identify 2–4 primary questions of interest to you about how you spend your time. Your questions should address something meaningful to you, but **be sure you identify questions you feel comfortable sharing with me** (your reports will not be shared with the class) and that they are feasible for the 14-day data collection period.

Feel free to expand upon the basic question of "Where does my time go?" or "How do I spend my time?" or explore a variation of it. Some other ideas include but are not limited to:

- Document *intended time* doing things (e.g. studying, sleeping) versus *actual time* doing those things, and compare results
- Document time spent on each course, and/or time spent on different parts of a course (e.g. in class, reading, homework, etc.)
- Document time spent on school vs. work vs. leisure vs. rest, etc.
- If you already use a calendar app as a way to keep track of your schedule, you could compare how your time was spent last year at this time versus how your time is being spent this semester.

You'll also need to identify appropriate visuals and at least one table that you'll use in your analysis. The visuals and table must convey different information. All of this information will be collected via a proposal for the project. You will need to fill in the project proposal template, compile to .pdf and then submit the .pdf on Gradescope for me to review and approve.

You should also go ahead and set up a calendar-query folder in your repo to put project materials, such as the proposal template, your proposal, the report template, your report, your data, etc. If you've pulled the calendar-query folder from the class repo, that should be as easy as copying it into your repo.

2. Collect data

Once your proposal is approved, for 14 days (or more!), track your time in a calendar application. I recommend Google calendar, but other electronic calendars are acceptable as long as you can download an `ics` file format, a universal calendar format used by several email and calendar programs.

If you want to incorporate other sources of data, I cannot guarantee that I (or the SDS Fellows) can provide support, so questions related to other sources should be supplemental.

Fill in blocks of time on your calendar, and mark an entry with the activity you were performing: sleeping, studying, eating, exercising, socializing, etc. How you fill in and categorize your blocks of time should depend on what your questions of interest are.

The default fields in Google calendar that can be read in easily are the activity name (Summary) and description. Unfortunately, grabbing "location" is proving to be more difficult than in the past, but you can put multiple pieces of information in the description and break them apart later.

Tips

- If you already use a calendar app, you should create a new calendar (or new calendars) within the app dedicated to this activity. That way, your pre-existing calendars will be kept private.
- Color coding of events is lost when exporting the data into the `ics` file, so *do not* rely on color-coding your calendar to give you information for this assignment. Your labels need to do it.
- Try to be consistent about how you use calendar fields.
- If you have multiple sources of data or are keeping track of your project data across multiple calendars (thus multiple `ics` files), I recommend creating a *data* subfolder in your calendar-query folder in your repo and placing all of your data files there.
- You'll want to iterate between collecting data and wrangling (next step) to identify problems with data collection as early as possible.
- Possible fields of interest include the summary (name of the event), and description fields. It is possible to include multiple pieces of information in one field by having a consistent system, such as entering something as `class-### activity` in the description, which could then be split on the space.
- IMPORTANT - Recurring events are NOT detected with the current read-in commands available. Recurring events should be entered individually for the purposes of this project.

3. Export, import, wrangle, and analyze!

Over the 14 days (or more!), regularly export your calendar data to the `ics` file format. This should take less than 5 minutes (e.g., see instructions for exporting one or more Google calendars).

Import the `ics` file into R as a dataframe using the **ical** package. You'll also need **tidyverse** and **lubridate** packages to accomplish this task. Sample code will be provided in the report template.

Create at least **two visualizations and one summary table** which convey distinct findings, wrangling the data as necessary along the way. Remember, you should be iterating between collecting data and wrangling to catch any problems as early as possible.

Again, the visuals and table should present different information. (If they display the same information, why have 2 versions of it?) Also, your table should *not* be a display of your raw data—instead, think about what summary statistics and/or numeric comparisons might be informative or useful to share. All visualizations should be constructed using ggplot. This is the standard graphics package for our class. In the rare event that you cannot make a particular plot in ggplot, consult with me.

For nicer tables, check out:

- the `kable()` function in the **knitr** package along with the **kableExtra** package (e.g., `kable_styling()`).
- the **xtable** package

I expect we will be able to use part of one class to work on wrangling our calendar data, but to take full advantage of that class time, it would be helpful if you attempted some wrangling prior to that class.

Do not go overboard with visualizations and tables. My suggestion (what I'm expecting, and the minimum you should have) is two visualizations and one table. If you need a third visualization or a second table to illustrate something, that's fine, but if you start needing many more, that's beyond the scope of the project.

4. Draw conclusions and communicate results.

What insights do you glean from your analysis? Write a report introducing your questions of interest, explaining what you found, and reflecting on the answers to your questions posed.

Your report should follow the provided report template and should answer the following questions:

- Describe the questions of interest: what are they and why are they important or useful to you?
- Briefly describe your data collection process, including how you defined variables of interest (include the levels/categories of categorical variables and the units for quantitative variables).
- In connection with the questions you posed, describe what information is conveyed through each visualization or table (avoid describing visualization choices).
- With your initial questions in mind, briefly summarize what you learned about how you spent your time. What are the big take-aways for you? If relevant, how does this affect how you might spend your time going forward?

5. Reflect

Finally, write a reflection on this experience (included at end of the report template). Particular questions to reflect on include...

- What difficulties in the data collection and analysis process did you encounter? Identify two of your main hurdles in gathering accurate data.
- What implications do those difficulties have for future data collection and/or analysis projects if you were to repeat this project?
- Did you have enough data? How much data do you think you'd need to collect in order to satisfactorily answer your question(s) of interest? Would it be hard to collect that data? Why or why not?
- As someone who provides data, what expectations do you have when you give your data (e.g. to Facebook, Google, MapMyRun, etc.)?

- As someone who analyzes others' data, what ethical responsibilities do you have?

The reflection is contained at the end of the report template. The final .pdf of the report will be submitted to Gradescope for me to review. All other project components - your .Qmd(s), your data, etc. will be contained in a folder in your git repo that will also be reviewed. The aim is to make sure your project is reproducible.

Timeline

Activity	**Timeline**
Proposal due	Monday, February 5th by midnight to Gradescope
Begin data collection	Wednesday, February 7th or upon approval
Export and wrangle data	Do this regularly throughout the 2-week period
Finish data collection	Tuesday, February 20th (or 2 weeks from approval)
Final report due	Wednesday, February 28th by midnight to Gradescope
All .Qmd(s), data, etc. files due	Wednesday, February 28th by midnight to your private repo

References

Thank you to prior Stat 231 instructors Correia and Bailey for sharing materials related to this project including example templates and the assignment. Thank you to Albert Kim (Smith College) and Johanna Hardin (Pomona College) for the Google Calendar assignment (see the set up for their courses here and here). They were inspired by an episode of the *Not So Standard Deviations* podcast titled "Compromised Shoe Situation", in which the hosts (Roger Peng and Hilary Parker) discuss a data science design challenge on getting to work on time.