Assignment 4

No changes in assignment description but some clarification added on Mar 26th.

You have two different options for the assignment. One is using Chris's exercise data --

This assignment is a real-world one, and Professor Chris Brooks is your client! In short, he started increasing his exercise over the summer of 2019 and started collecting data on what he was doing. Throughout the summer he bought a variety of devices (heart rate monitor, watch, bicycle, etc.), and began publishing this data to the social sharing site strava.

The other option is to use a survey data composed of 13 questions. Here is the data description by Grant:

"The survey has a bit of a twist, for each question you will also be asked to predict your classmates' answers. One of the things you can do is then see whose predictions are the most accurate. Or see how predictions differed between those who responded differently."

In both cases, your job in this assignment is to explore the data dump and say something interesting about it. You will be graded on:

1. (20%) Are you making a compelling computational narrative, judged in part by Rule et al's ten rules for computational analyses?
    a. You don't need to follow all of the rules all of the time, but you must explicitly indicate at the header of each notebook which rules you adhered to and what the evidence was.
2. (45%) Have you demonstrated that you have a solid grasp of at least three of the basic visual analysis techniques in this class (**scatter, box, line, violin, histograms, heatmaps, probability plots, treemaps, sploms**) and that they were appropriate for the analysis/data you were investigating?
    a. You get equal grades for each plot type (15% each), and grades for a given plot will be broken down into three equal categories (5% each):
        i. The mechanics of generating a reasonable plot from the data you are working with.
        ii. The justification for the plot and the insight as a result, as described by your computational narrative.
        iii. Making the plot rock visually, by embedding advanced features ranging from the aesthetic (color, form) to the informational (callouts, annotations).
3. (15%) Have you demonstrated that you have a solid grasp of at least one of the more advanced visual analysis techniques in this class **(Don't use any visualizations listed as basic plots above. You can explore a new visualization technique which the lecture didn't teach you, or you can even come up with a combination of multiple types of plots to generate an advanced plot)** and that it was appropriate for the analysis/data you were investigating? The grading rubric is the same as the basic plots. You may use other advanced plots with permission in this category (ask first to ensure they seem reasonably advanced).
4. (20%) Are you able to provide an interesting and defensible analysis that helps Professor Brooks understand what this data means in the context of his activities? If your data science discovery will make the client happy then this part of the overall grade tilts up towards 20%. If there are obvious things you should have looked at then it tilts down towards 0%.

Note 1: You should not redistribute this data. Thanks :)

Note 2: Your client is genuinely excited to read what you put together! Please discuss with one another about the kinds of data that are in the data files but do not share code.

Note 3: The data can be found in your Coursera Labs image as the file "strava.csv".

Note 4: You can use one or more notebooks to answer this project as you see fit, there is no requirement for multiple notebooks. As described in grading rubric #1, it is expected that you will provide a narrative of how you adhered to or interpreted several of Rule et al's heuristics for computational narratives. Normally this would not actually exist in a narrative document, this is solely to demonstrate your ability to internalize these rules and reflect on your own work through this lens. While there are no hard limits on the number of rules you should address, I expect at least three rules would be able to be discussed for a notebook of this size, and that discussion and evidence of how you aligned with those rules would be on the order of 1-2 paragraphs per rule.

Frequently Asked Questions

1. **Q. What are the units of the data?**
   A. The data are in a variety of different units. A previous student noted the following units:
   Cadence: rpm
   Ground time: milliseconds
   Vertical oscillation: centimeters
   Distance, Altitude, and Enhanced Altitude: meters
   Longitude and Latitude: semicircles (radians)
   Air and Form Power: watts
   Leg Spring Stiffness: kN/m
   Speed: m/s
2. **Q. How can I explore the units more?**
   A. Here's some code a student wrote last year using the FitFile package:

```
from fitparse import FitFile
import pandas as pd
import numpy as np

# Get all data messages that are of type record
stryd = FitFile('withstryd.fit').get_messages('record')
my_data = FitFile('my_data.fit').get_messages('record')

def get_data_plus_units(messages):
    data_record = []
    for record in messages:
        # Go through all the data entries in this record
        record_datum = {}
        for record_data in record:
            # Print the records name and value (and units if it has any)
            if record_data.units:
                record_datum[record_data.name] = str(record_data.value) + ' ' + record_data.units
            else:
                record_datum[record_data.name] = str(record_data.value) + ' unknown'
        data_record.append(record_datum)
    return data_record

df = pd.DataFrame(get_data_plus_units(stryd))
df2 = pd.DataFrame(get_data_plus_units(my_data))
df.head(10).transpose()
```

3. **Q. Where can I find the precise meaning of these columns?**
   A. The precise meaning of these columns is not defined. This is an "authentically messy and unclear" dataset for you to explore and try and understand, you are welcome to ask me or discuss with peers based on googling around and the details I put in the assignment. I'm not intentionally hiding information I know!
4. **Q. How can I tell the difference between cycling/running activities?**
   A. I think some broad understanding of the difference between cycling and running (e.g. speed which can be attained, distances) is probably the best way to start exploring this.

5. **Q. There are fields like cadence/Cadence or altitude/enhanced_altitude. What's up?**
   A. These are *likely* from different devices, but that's a best guess. Some of the devices used were stryd, garmin forerunner 245 music, garmin cadence sensor, garmin speed sensor. The last two were used on bike, the first one was used running, and the forerunner was used during both (to receive data).
6. **Q. Where are you able to find the 10 principles to design good figures?**
   A. You can watch the design principles at https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833
7. **Q. May we extract the information from the .fit files in bonus instead of using the strava.csv file?**
   A. Yes, keep in mind that the fit files in the bonus assignment are a superset of the data in strava.csv (they cover a longer time period).
8. **Q. Are we permitted to use any graphing library for Assignment 4?**
   A. Yes.
9. **Q. What is an individual exercise versus a portion of an exercise?**
   A. This is up to your interpretation! There are at least some activities where I ran then stopped for a while then ran again. Some days have multiple runs in them, for instance. You tell me what is reasonable as far as splitting up this data as you explore it.
10. **Q. Can you explain #4 on the grading. What do you mean by "interesting and defensible analysis that helps you understand what this data means in the context of your activities?"**
    A. This is the most subjective portion of the assignment. I want to see you give me a summary of what you discovered that demonstrates your knowledge of the issues discussed in the course. It's vague as there are many ways to provide a reasonable explanation -- I want to see writing at a graduate level which is reasonable for the task/discovery/methods. I know that isn't very clear but my goal isn't to trick or deduct points without reason, I just want to see what insights you have actually hunted down (and your methods).