Contents

Assignment 3

The purpose of this assignment is two-fold: (a) to have you engaged in independent learning with respect to a novel visual exploration method, a novel python-based information visualization toolkit, or both, and (b) to enhance our knowledge overall in the class (including mine!) by engaging in peer instruction. More specifically, I want you to write a tutorial demonstration which shows a new visualization technique (e.g. pareto charts) and/or a new visualization toolkit (e.g. bokeh, altair, seaborn, plot.ly). Feel free to do both a new technique and a new toolkit.

The purpose of using a data visualization is to tell a story, and the purpose of writing up a demonstration is to teach someone else about what you have learned so that they can use it, too. Think of this as a worked example, something like a medium blog post that you have written up and are posting so that other data scientists can learn about the potential uses of this style of data exploration and visualization. You should be concise in your wording and clear in your demonstration.

You will need to include the write up and demonstrate in a new notebook which you will be submitting to the staff. Your notebook should be primarily narrative in nature, broken up by code cells that demonstrate the work that you're demonstrating. You should have the following main sections in your notebook (but feel free to use other subheadings as appropriate):

- Visualization Technique (25%)
    - A narrative description of the visualization you are planning to use, describing how it works
    - A discussion of in which circumstances this visualization should and should not be used (what is is close to? What else could you consider? How does it relate to specific aspects of data?
- Visualization Library (25%)
    - The library you are going to use, and a background on why the library is good for this visualization. Who created it? Is it open source? How do you install it?
    - A discussion of the general approach and limitations of this library. Is it declarative or procedural? Does it integrate with Jupyter? Why you decided to use this library (especially if there are other options)?
- Demonstration (50%)
    - The dataset you picked and instructions for cleaning the dataset. You should pick a suitable dataset to demonstrate the technique, toolkit, and problem you are facing.
    - The quality of your demonstration. First demonstrate the basics of this approach, then show a few of the edges of how the library might be used for other cases. This is the "meat" of the assignment.

Assignments should be submitted as both your source code as an .zip file of your notebook (.ipynb) and other assets (e.g. .csv) as well as a .pdf of this notebook which we can view in a browser.

Frequently Asked Questions

1. Q. This sounds daunting, can I have an example?
   A. Yes! Anthony Giove, a student from the Fall 2019 cohort, has provided his example for you in the example_assignment directory in your workspace. This example is a stellar one, and is above and beyond what you need to do for this assignment. In particular, you don't need to find such an extensive set of data, nor do you have to go all the way into advanced features such as interactivity. Yet, I think many of you will find this motivating and I look forward to seeing what you came up with!
2. Q. This assignment is daunting! How big does it have to be?

   A. Think a couple of pages, one compelling visual, or a couple of less compelling visuals. Make sure the question you ask is well aligned with the technique/library you are using.
3. Q. So the visual and technique is the important part, where the insights are less so?

A. Yes, this is a good interpretation. I care less about the outcome and more about your learning of a technique or library.

4. Q. Can I use tableau?

A. No; it needs to be a python toolkit.

5. Q. What do you mean by declarative or procedural?

A. Conceptually, Declarative programming is where you say what you want without having to say how to do it. With procedural programming, you have to specify exact steps to get the result.

A more full discussion of declarative or procedural should be included in the lectures, however the assignment has been updated (see changes above) to clarify that these were to be example topics you might cover for each of the prompts given.

6. Q. Is a new visualization technique absolutely necessary? Could I use a new visualization package while using the visualization technique learned during the lecture?

A. Yes! This assignment requires you to write a tutorial which shows a new visualization technique and/or a new visualization toolkit or package. This means you can use an old visualization technique while using a new visualization toolkit.

7. Q. Can I have more examples about what visualization package we can use for assignment 3?

A. Yes! Apart from the packages mentioned above, we can use plotly, bokeh, dataviz, vega, altair, folium, geopandas, earthpy, etc. Those are wonderful packages to explore.

8. Q. I have some data that I want to explore. How shall I know which visualization techniques and visualization toolkits are suitable for such analysis?

A. While the answer varies depending on the nature of your dataset and your visualization approach, a good way to research your problem is through looking at the gallery page for a couple of different visualization toolkits. If you look closely you will find at least some useful graphs to represent your data to the audience.

9. Q. Do we need to provide references for the narrative information relevant to the visualization technique or the visualization library? If yes, what citation style is recommended?

A. A URL or APA citation style would be fine.

Bonus Assignment

This assignment is hard. It's meant to be a way for you to distinguish yourself in this class, and to show off your skills in python, data manipulation, and visualization. It's also the only way you can achieve an A+ in the course, and is worth up to 5% of your final grade, though it's expected that of the few who attempt the assignment most will only get partial grades. For this assignment, there is no help provided by the teaching staff, because of bandwidth constraints, but you are welcome to discuss with your peers and share information on how to accomplish the tasks (but no code sharing, please!). Impress me!

In preparing the MADS curriculum I had a discussion with a colleague about the value of dashboards and information visualization. Dashboards are now ubiquitous in any consumer-facing analytics system, yet knowledge of their effectiveness, utility, or even ideas towards design patterns for building dashboards are limited. In this bonus assignment, you will explore the creation of a dashboard for the bulk of fitness data I've put in your coursera resources/bonus folder. This folder is made up of ~200 files which are in the Flexible and Interoperable Data Transfer, and include temporal, geographic, and sensor-based measurement data related to activity.

As you know, I already have a couple of dashboards at my disposal, and I've included a copy of my strava dashboard for a single activity, and my garmin dashboard for a single activity as appendices in this assignment. What I want you to do is to design me a new dashboard, all within the Jupyter notebook environment. I have ~~four~~ three requirements for this dashboard:

1. It should be based on one or more well articulated design principles. I expect a short description of how you designed the dashboard to align with some design principles.
2. You must use some Jupyter widgets to add interactivity. You might find the following links useful:
    a. https://ipywidgets.readthedocs.io/en/latest/examples/Widget%20List.html#
    b. https://towardsdatascience.com/bring-your-jupyter-notebook-to-life-with-interactive-widgets-bc12e03f0916
    c. https://medium.com/plotly/introducing-jupyterdash-811f1f57c02e
3. You must take advantage of the three dimensions of the data - temporal, geographical, and analytical - in your dashboard.

Frequently Asked Questions

1. **Q. How do I work with FIT data?**
   A. We've installed the python-fitparse library for you to manipulate the data. The following code example will parse all of the datafiles and print out the mean heart rate and time, you might find it handy to start with this.

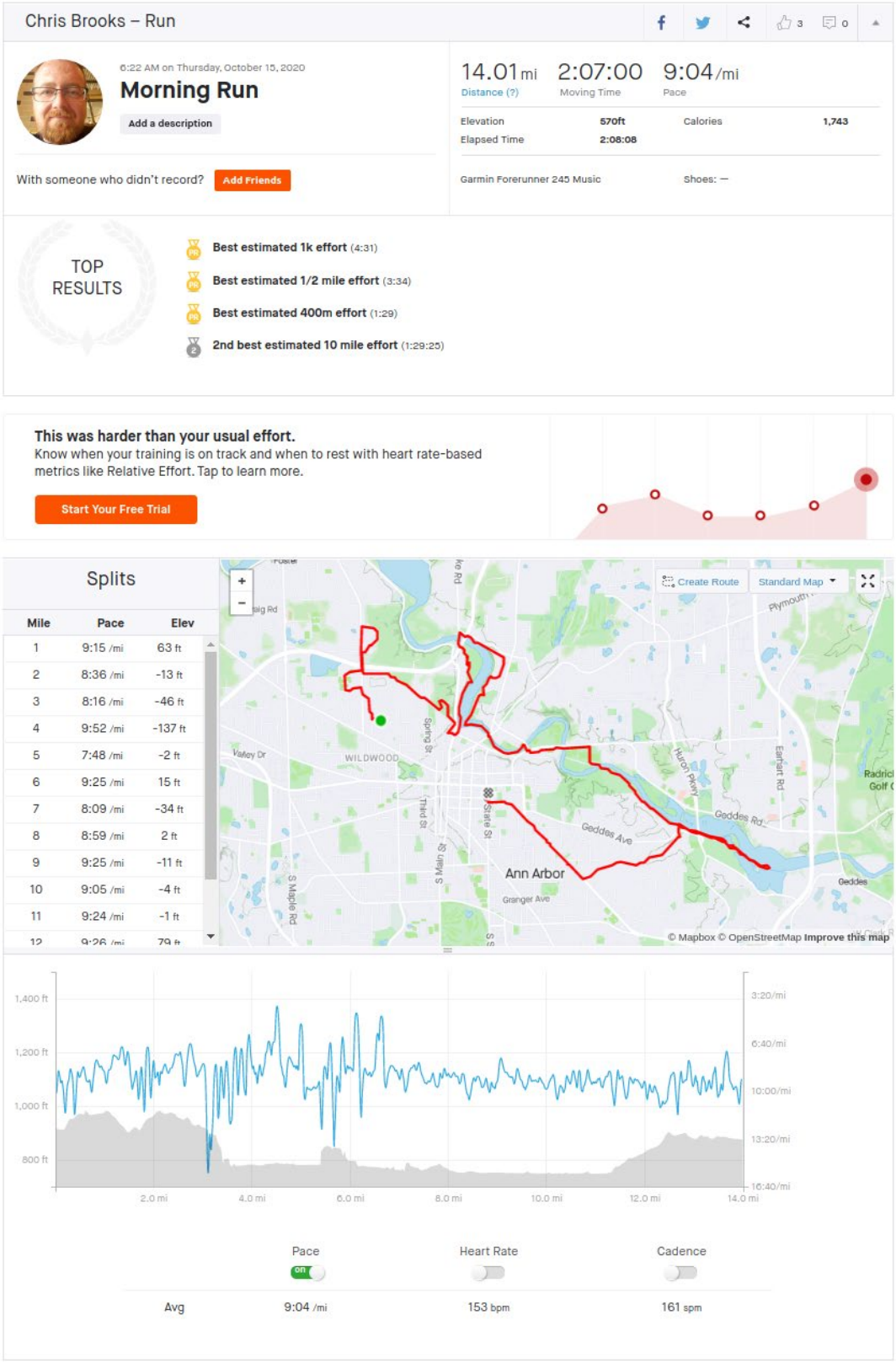
```
import pandas as pd
import numpy as np
from fitparse import FitFile


datafiles=!ls bonus/*.fit


for datafile in datafiles:
    with FitFile(open(datafile, 'rb')) as fitfile:
        df=pd.DataFrame([record.get_values() for record in fitfile.get_messages('record')])
```

```
if "timestamp" in df and "heart_rate" in df and len(df['heart_rate'].dropna())>0:
    print(f"Mean heart rate for activity on {df['timestamp'].iloc[0]} was {np.nanmean(df['heart_rate'])}.")
```

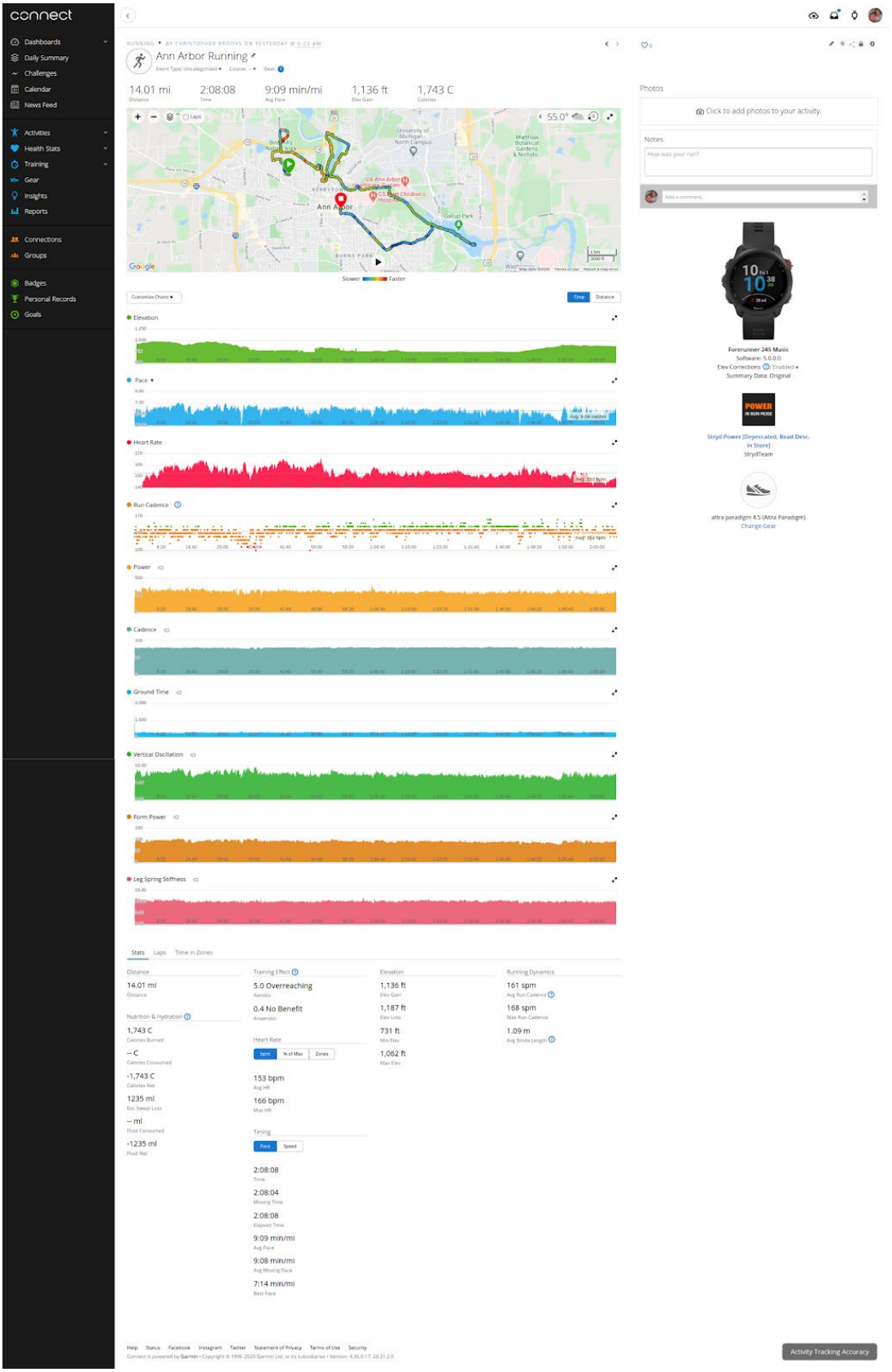
2. Q. How do I submit the bonus?

A. Email it directly to Chris at brooksch@umich.edu by the deadline, **which is November 23rd at 11:59 PM EST**. Please include your notebook, datafiles (other than the fit files), and a PDF or something easy for Chris to read!

Appendix 1: Strava Dashboard Example



Appendix 2: Garmin Connect Dashboard Example



Assignment 4

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. Your job in this assignment is to explore his strava 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?

. 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 (time series, 3d, geographic/mapping, spatial) 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).