MADA Project Part 1

Joe Martin

9/12/2021

## Run Data: 2013 - Present

#### Question 1

**Briefly describe what the data is, how it was collected, where you will get (or got) it from. How many observations do you have, what was measured? Anything else important to report about the data?**

The data I intend to use for my project is my personal running data. I started distance running in 2012 just before I got my very first smartphone. At the beginning of 2013, I started to track all of my runs in the Nike Run Club app. This was my primary way of tracking my runs. Toward the end of 2014, I started trying various other apps and mostly got the same data out of them. Apps included MapMyRun and Strava, and possibly a few more. Much of my data from 2015 and 2016 is missing and I’m unsure how to, or if I could ever, recover it. In 2017 I mostly used MapMyRun. In 2018, I moved back to Nike Running and had a full month in 2020 when I used Strava exclusively. In April of 2020, I bought my first smartwatch (a Garmin Forerunner 245), which I’ve been using to track all of my training (running, swimming, and cycling).

I estimate that I have about 1,000 observations for running. In 2020 when we went into lockdown for COVID-19, I began moving all of my data from my various apps and accounts into a Google sheet. I mainly did this manually. I enter my data into another spreadsheet manually for my most recent runs, which I primarily use as my training journal. In this journal, I record imprecise numbers – average pace, mile/lap splits, the shoes I wore, weather, time of day/date, total distance, location of my run, distance type (short, medium, or long), and workout type (recovery, easy, speed, long distance). Along with this, I’ll also take notes about my run, like how I feel, thoughts about how my training is going, whether I’m feeling any pain or note any injuries, etc. With my Garmin watch, I have a Garmin Connect account, which automatically stores my data to the cloud. When I export data related to my activities (running, swimming, and cycling all come together), Garmin gives me a .csv file with 36 variables, although some are empty or have 0 for all values.

Outside of this data, I also have access to data about my sleep and resting heart rate from my Garmin watch. Sleep data includes how long I slept, how long I spent in light sleep, deep sleep, REM sleep, and my respiration during sleep. Resting heart rate data is included below and shows my resting heart rate for each day. I’m interested in using this data and even reviewing activities outside of running (swimming and cycling) to see if they affect my performance, although I’m also hesitant to start with such a broad scope.

Going through my data, I have noticed that most of my variables are currently stored as character data. This will be one of the first things I will need to fix to make the most out of my data.

## Question 2

**At this stage you are not required to already have and show the data, but if you do, even better. Then add a few lines of code which load the data and using some of the commands you learned about, provide summary descriptions of the data.**

Data included in this section contains my run logs from 2013 - 2021, as well as my resting heart rate from April 15, 2020 - present. Data I have available for use, but not yet ready to share includes my detailed post-run notes, sleep data, and other random calculations my Garmin watch does for me (VO2 Max, Pulse OX, stress, etc.). I’m leaving these out initially because, while I’m interested in using them, I am also concerned they will greatly broaden the scope and complexity of this project.

#### Historical Data 2013 - 2018

summary(df1318)

## day date weight start\_time   
## Length:302 Length:302 Mode:logical Length:302   
## Class :character Class :character NA's:302 Class :character   
## Mode :character Mode :character Mode :character   
## run\_type distance\_type weather location   
## Mode:logical Length:302 Length:302 Length:302   
## NA's:302 Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## run\_shoes distance total\_time average\_pace   
## Length:302 Length:302 Length:302 Length:302   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
## splits notes   
## Length:302 Length:302   
## Class :character Class :character   
## Mode :character Mode :character

head(df1318)

## # A tibble: 6 x 14  
## day date weight start\_time run\_type distance\_type weather location  
## <chr> <chr> <lgl> <chr> <lgl> <chr> <chr> <chr>   
## 1 Sunday 1/20/2013 NA 10:06 AM NA Short Distance 55,clear Bonterr~  
## 2 Thursday 2/21/2013 NA 3:34 PM NA Short Distance <NA> Bonterr~  
## 3 Saturday 4/6/2013 NA 10:33 AM NA Short Distance 59,clear Bonterr~  
## 4 Monday 4/8/2013 NA 4:40 PM NA Short Distance 75,clou~ Bonterr~  
## 5 Tuesday 4/9/2013 NA 4:37 PM NA Short Distance 79,part~ Bonterr~  
## 6 Thursday 4/11/2013 NA 10:11 AM NA Short Distance 70,clou~ Bonterr~  
## # ... with 6 more variables: run\_shoes <chr>, distance <chr>, total\_time <chr>,  
## # average\_pace <chr>, splits <chr>, notes <chr>

#### Data Combined 2019 - 2021

summary(df1)

## weekDay date Activity Type dateTime   
## Length:974 Length:974 Length:974 Length:974   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Favorite Title Distance Calories   
## Mode :logical Length:974 Length:974 Min. : 79.0   
## FALSE:372 Class :character Class :character 1st Qu.: 436.0   
## NA's :602 Mode :character Mode :character Median : 638.0   
## Mean : 750.9   
## 3rd Qu.: 951.0   
## Max. :2089.0   
## NA's :601   
## Time Avg HR Max HR Aerobic TE   
## Length:974 Min. :134.0 Min. :149.0 Min. :1.000   
## Class :character 1st Qu.:157.0 1st Qu.:173.0 1st Qu.:3.100   
## Mode :character Median :163.0 Median :179.0 Median :3.700   
## Mean :160.9 Mean :176.9 Mean :3.591   
## 3rd Qu.:168.0 3rd Qu.:183.0 3rd Qu.:4.100   
## Max. :182.0 Max. :199.0 Max. :5.000   
## NA's :601 NA's :601 NA's :601   
## Avg Run Cadence Max Run Cadence Avg Pace Best Pace   
## Min. :140.0 Min. :152 Length:974 Length:974   
## 1st Qu.:147.0 1st Qu.:163 Class :character Class :character   
## Median :149.0 Median :169 Mode :character Mode :character   
## Mean :149.7 Mean :173   
## 3rd Qu.:152.0 3rd Qu.:180   
## Max. :162.0 Max. :245   
## NA's :601 NA's :601   
## Elev Gain Elev Loss Avg Stride Length Avg Vertical Ratio  
## Length:974 Length:974 Min. : 1.010 Min. :0   
## Class :character Class :character 1st Qu.: 1.300 1st Qu.:0   
## Mode :character Mode :character Median : 1.380 Median :0   
## Mean : 1.825 Mean :0   
## 3rd Qu.: 1.458 3rd Qu.:0   
## Max. :175.000 Max. :0   
## NA's :600 NA's :601   
## Avg Vertical Oscillation Training Stress Score® Grit Flow   
## Min. :0 Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :0 Max. :0   
## NA's :601 NA's :601 NA's :601 NA's :601   
## Climb Time Bottom Time Min Temp Surface Interval   
## Length:974 Length:974 Min. :0 Length:974   
## Class :character Class :character 1st Qu.:0 Class :character   
## Mode :character Mode :character Median :0 Mode :character   
## Mean :0   
## 3rd Qu.:0   
## Max. :0   
## NA's :626   
## Decompression Best Lap Time Number of Laps Max Temp   
## Length:974 Length:974 Min. : 1.00 Min. :0   
## Class :character Class1:hms 1st Qu.: 4.00 1st Qu.:0   
## Mode :character Class2:difftime Median : 7.00 Median :0   
## Mode :numeric Mean : 7.71 Mean :0   
## 3rd Qu.: 9.00 3rd Qu.:0   
## Max. :23.00 Max. :0   
## NA's :626 NA's :626   
## Run Type Distance Type Weather Location   
## Length:974 Length:974 Length:974 Length:974   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Shoes   
## Length:974   
## Class :character   
## Mode :character   
##   
##   
##   
##

head(df1)

## # A tibble: 6 x 37  
## weekDay date `Activity Type` dateTime Favorite Title Distance Calories Time   
## <chr> <chr> <chr> <chr> <lgl> <chr> <chr> <dbl> <chr>  
## 1 Tuesday 1/1/~ <NA> 2019-01~ NA <NA> 3.04 NA 00:2~  
## 2 Wednesd~ 1/2/~ <NA> <NA> NA <NA> <NA> NA <NA>   
## 3 Thursday 1/3/~ <NA> <NA> NA <NA> <NA> NA <NA>   
## 4 Friday 1/4/~ <NA> <NA> NA <NA> <NA> NA <NA>   
## 5 Saturday 1/5/~ <NA> <NA> NA <NA> <NA> NA <NA>   
## 6 Sunday 1/6/~ <NA> 2019-01~ NA <NA> 3.01 NA 00:2~  
## # ... with 28 more variables: Avg HR <dbl>, Max HR <dbl>, Aerobic TE <dbl>,  
## # Avg Run Cadence <dbl>, Max Run Cadence <dbl>, Avg Pace <chr>,  
## # Best Pace <chr>, Elev Gain <chr>, Elev Loss <chr>, Avg Stride Length <dbl>,  
## # Avg Vertical Ratio <dbl>, Avg Vertical Oscillation <dbl>,  
## # Training Stress Score® <dbl>, Grit <dbl>, Flow <dbl>, Climb Time <chr>,  
## # Bottom Time <chr>, Min Temp <dbl>, Surface Interval <chr>,  
## # Decompression <chr>, Best Lap Time <time>, Number of Laps <dbl>, ...

#### Resting Heart Rate 4/15/21 - Present

rhr\_data <- here::here("data", "dailyRHR.csv")  
  
rhr <- read\_csv(rhr\_data)

## Rows: 518 Columns: 3

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): date, weekday  
## dbl (1): rhr(bpm)

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

summary(rhr)

## date weekday rhr(bpm)   
## Length:518 Length:518 Min. :41.00   
## Class :character Class :character 1st Qu.:45.00   
## Mode :character Mode :character Median :47.00   
## Mean :47.28   
## 3rd Qu.:48.00   
## Max. :74.00   
## NA's :3

head(rhr)

## # A tibble: 6 x 3  
## date weekday `rhr(bpm)`  
## <chr> <chr> <dbl>  
## 1 4/15/2020 Wed 48  
## 2 4/16/2020 Thu 48  
## 3 4/17/2020 Fri 49  
## 4 4/18/2020 Sat 43  
## 5 4/19/2020 Sun 41  
## 6 4/20/2020 Mon 42

## Question 3

**Explain the question you want to answer using the data. What will be your outcome(s) of interest (if any)? What (if any) specific predictors will you focus on? What relations/patterns are you looking for in the data?**

Through this project, I’m interested in understanding how factors of my training program affect my performance. I would like to see if there are any significant relationships between the frequency or intensity of my training and performance. My outcomes of interest would be finding whether there are significant patterns in my pace, active heart rate, or resting heart rate. I’m also interested in seeing if I can measure qualitative outcomes through the notes I take after each run. Some of the patterns I would look for in my data would include the relationship between a positive performance outcome and variables like shoes, number of speed workouts before a race or a long run, or weekly volume (total weekly mileage or time exercising).

**As much as you know, suggest how you will analyze it. At this stage in the course, we haven’t covered analysis approaches yet, so you can keep things vague and non-technical here.**

In order to search for relationships and patterns in my data, I will first need to identify specific events - like races or long runs - and compare weeks of training against each other. For example, I do one long distance run just about every week. I would be interested in understanding if my pace decreases on a long run if I had more active training days leading up to this run. I would want to see if my heart rate during similar runs (for example, long runs on the same route) decreases over time or increases after a break in training.

The other way I may go about analyzing patterns in my data would be by picking specific distances described the same way (easy, steady, tempo). For example, a common workout for me would be an up-tempo 10k (6.22 miles) or an easy 8-miler.

Finally, although I have few races in my data set, I do have a good understanding of my training program, which I follow for weeks in advance of a race. For some races, I have a detailed log of this program. I could measure my fitness ahead of a training program and compare it to my fitness just before or at race time. Measurement of fitness would include average pace, heart rate, and perhaps even ratings based on positive or negative notes taken about training sessions.

## Question 4

**You are allowed, but not yet required, to provide background information for the question you plan to answer. For instance you can describe why it’s an interesting question, who else has done similar analyses, how your analysis will be new/different, etc. Similar to what you read in an introduction to a research paper. For the final report, you’ll need these parts. For part 1, they are not required, but you are welcome to already scribble down some of that.**

#### Background

Coaches and personal trainers are expensive to hire. Apps provide users with some basic stats about their runs, but users typically have to purchase a subscription to get useful insights into training. Further, apps often use algorithms that do not produce practical insights. For example, Garmin provides users with a Performance Condition measurement. Garmin devices display this measurement to users during a workout to tell them how their performance compares to their average fitness level. Using pace, heart rate, and heart rate variability data, Garmin displays a score from -20 (Poor) to values greater than 10 (Excellent).

This is just one example of a measurement which ignores natural variability in training sessions and directly contradicts training programs established over decades by expert coaches and researchers, like Pete Pfitzinger and Jack Daniels, PhD. These expert coaches outline training plans which have a general pattern for long distance races, from 5,000 m races, all the way to marathons. This pattern includes two general aerobic runs per week (about 1:00 per mile slower than race pace), one or two speed sessions (lactate threshold runs at race pace, tempo runs, or track intervals), one or two recovery runs (about 2:00 per mile slower than race pace or heart rate in or below Zone 3), and a long distance run. These programs build in rest days, as well, which athletes are supposed to use for recovery or cross training.

Aggregate fitness data used by companies like Garmin and Polar is useful developing insights into human performance outcomes for the average person, like in “Human running performance from real-world big data” by Thorsten Emig & Jussi Peltonen. Due to the variability in quality distance training programs, however, aggregate measures and seemingly random algorithms are not useful in helping individual athletes understand whether they’re training well.