Marathon Training Plan Data Cleaning

Steve Newns

January 18, 2017

# *Initial Data Inspection*

## Strava Data File

First, I install packages

#install.packages("stringr") - for manipulating string values  
#install.packages("tidyr") - data cleansing  
#install.packages("plyr") - for renaming columns

Next, I load in the CSV file containing the Strava data file created by and downloaded from **VeloViewer.com** and check out the dataset.

setwd("C:/Users/snewns/Dropbox/RunningAnalysis/Data") #MOVE TO C FOR FINAL DRAFT  
strava <- read.csv("strava.csv")  
dim(strava)

## [1] 201 85

head(strava)

## Activity.Score When Type Gear  
## 1 50 11/20/2016 7:00 Run Saucony Fastwitch Reds  
## 2 0 11/19/2016 16:48 Run Hoka One One Clifton Reds  
## 3 0 11/18/2016 17:22 Run Hoka One One Clifton Reds  
## 4 0 11/17/2016 6:01 Run Hoka One One Clifton Reds  
## 5 0 11/16/2016 12:07 Run Hoka One One Clifton Reds  
## 6 0 11/15/2016 5:55 Run Saucony Fastwitch Reds  
## Name Dist.mi Elv.ft Elapsed.Time  
## 1 Oh My God So Freakin Windy Philly Marathon 43119.2 136.0 11443  
## 2 Recovery Run 4832.0 37.3 1571  
## 3 Recovery Run 8050.4 64.8 2616  
## 4 GA Run 11270.1 50.8 3159  
## 5 Recovery Run 8058.0 36.8 2404  
## 6 GA Run 12880.4 91.3 3638  
## Moving.Time Speed.mph Max.Speed.mph Pace..mi Max.Pace..mi Pace..100yds  
## 1 11234 3.8383 10.1 419.2341 159.3411 23.82012  
## 2 1515 3.1894 3.8 504.5293 423.5150 28.66644  
## 3 2530 3.1820 4.2 505.8168 383.1848 28.73959  
## 4 3042 3.7048 5.5 434.3619 292.6109 24.67966  
## 5 2328 3.4613 4.5 464.9395 357.6284 26.41702  
## 6 3494 3.6864 5.6 436.6150 287.3806 24.80767  
## Max.Pace..100yds Pwr.W Weighted.Avg.Pwr.W Max.Pwr.W Power.Meter Cad  
## 1 9.053474 0 0 NA 0 93.1  
## 2 24.063350 0 0 NA 0 87.6  
## 3 21.771864 0 0 NA 0 87.3  
## 4 16.625621 0 0 NA 0 92.8  
## 5 20.319797 0 0 NA 0 90.2  
## 6 16.328441 0 0 NA 0 93.5  
## Heart Max.Heart Elev.Dist.ft.mi Elev.Time.ft.h W.HR Temp.Â.C Cal  
## 1 151.1 200 3.15 43.6 0 NA 4226.9  
## 2 123.8 144 7.72 88.6 0 NA 473.6  
## 3 132.7 148 8.05 92.2 0 NA 789.3  
## 4 144.9 194 4.51 60.1 0 NA 1103.5  
## 5 134.6 164 4.57 56.9 0 NA 790.1  
## 6 144.8 162 7.09 94.1 0 NA 1266.0  
## Energy.kJ Segs PRs Others Kudos Comments Com Trn Man Pri Dist.away.mi  
## 1 0 17 17 118 3 0 0 0 0 0 -1  
## 2 0 0 0 1 0 0 0 0 0 0 -1  
## 3 0 0 0 1 0 0 0 0 0 0 -1  
## 4 0 0 0 1 0 0 0 0 0 0 -1  
## 5 0 0 0 1 0 0 0 0 0 0 -1  
## 6 0 0 0 1 0 0 0 0 0 0 -1  
## Activity.Score.25. Segment.PRs..K.QOM.positions.  
## 1 50 NA  
## 2 0 NA  
## 3 0 NA  
## 4 0 NA  
## 5 0 NA  
## 6 0 NA  
## All.segments..personal.positions. Ride.Run.Type X400m X1.km X5.km X10.km  
## 1 NA Race 78 215 1159 2371  
## 2 NA 116 303 NA NA  
## 3 NA 114 286 1633 NA  
## 4 NA 94 253 1367 2792  
## 5 NA 107 282 1468 NA  
## 6 NA 94 244 1315 2767  
## X15.km X20.km X30.km X50.km X1.2.Mile X1.mile X2.mile X10.mile  
## 1 3650 4888 7536 NA 170 360 751 3933  
## 2 NA NA NA NA 241 495 1036 NA  
## 3 NA NA NA NA 230 466 981 NA  
## 4 NA NA NA NA 202 412 842 NA  
## 5 NA NA NA NA 223 452 935 NA  
## 6 NA NA NA NA 195 400 802 NA  
## Half.Marathon Marathon Suffer.Score Points.in.red X..in.red H.R.Zone.1  
## 1 5157 11207 0 0 0 0  
## 2 NA NA 0 0 0 0  
## 3 NA NA 0 0 0 0  
## 4 NA NA 0 0 0 0  
## 5 NA NA 0 0 0 0  
## 6 NA NA 0 0 0 0  
## H.R.Zone.2 H.R.Zone.3 H.R.Zone.4 H.R.Zone.5 Power.0W Power.0.50W  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## Power.50.100W Power.100.150W Power.150.200W Power.200.250W  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0  
## Power.250.300W Power.300.350W Power.350.400W Power.400.450W Power.450W.  
## 1 0 0 0 0 NA  
## 2 0 0 0 0 NA  
## 3 0 0 0 0 NA  
## 4 0 0 0 0 NA  
## 5 0 0 0 0 NA  
## 6 0 0 0 0 NA  
## Pace.Zone.1 Pace.Zone.2 Pace.Zone.3 Pace.Zone.4 Pace.Zone.5 Flagged  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## Description City State Activity.Id  
## 1 NA NA 780935487  
## 2 NA NA 780934297  
## 3 6\*100m strides NA NA 779226144  
## 4 NA NA 777917114  
## 5 NA NA 777267101  
## 6 2 mi @ MP (3-4) NA NA 776017804

Okay, so there's 201 activities in here, as well as a lot of columns (85 of them), with a majority of these columns probably not being useful for my purposes right now.

Now I want to explore the Garmin data as well before moving on the removing those unneccessary columns.

## Garmin Data Files

First, I load in the first of the CSV files containing the Garmin data created by and downloaded from **Garmin Connect** and check out the dataset.

setwd("C:/Users/snewns/Dropbox/RunningAnalysis/Data")  
garmin1 <- read.csv("garmin1.csv")  
dim(garmin1)

## [1] 21 16

head(garmin1)

## Activity.Name Start Time Distance  
## 1 Sea Isle City Running Wed, Aug 10, 2016 9:41 AM 23:31 3.00  
## 2 Sea Isle City Running Mon, Aug 8, 2016 7:31 AM 1:57:31 15.00  
## 3 Sea Isle City Running Sun, Aug 7, 2016 2:50 PM 30:04:00 4.01  
## 4 Ocean City Running Sat, Aug 6, 2016 6:18 AM 1:02:54 8.00  
## 5 Chalfont Running Fri, Aug 5, 2016 5:30 AM 1:30:35 12.00  
## 6 Chalfont Running Thu, Aug 4, 2016 5:11 AM 41:04:00 5.00  
## Elevation.Gain Avg.Speed.Avg.Pace. Max.Speed.Best.Pace. Avg.HR Max.HR  
## 1 -- 7:50 6:13 134 151  
## 2 13 7:50 6:12 142 158  
## 3 4 7:30 5:11 149 160  
## 4 -- 7:52 6:02 149 158  
## 5 459 7:33 5:05 148 158  
## 6 224 8:12 7:16 133 147  
## Calories sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF  
## 1 299 -- -- -- -- --  
## 2 1,545 -- -- -- -- --  
## 3 460 -- -- -- -- --  
## 4 898 -- -- -- -- --  
## 5 1,260 -- -- -- -- --  
## 6 545 -- -- -- -- --  
## Training.Effect  
## 1 --  
## 2 --  
## 3 --  
## 4 --  
## 5 --  
## 6 --

So there's 21 runs in here, with 16 variables. Let's look at another Garmin data file.

setwd("C:/Users/snewns/Dropbox/RunningAnalysis/Data")  
garmin6 <- read.csv("garmin6.csv")  
dim(garmin1)

## [1] 21 16

head(garmin6)

## Activity.Name Start Time Distance  
## 1 Philadelphia Running Sun, Nov 20, 2016 7:00 AM 3:07:15 26.27  
## 2 Chalfont Running Sat, Nov 19, 2016 4:48 PM 25:16:00 3.00  
## 3 Chalfont Running Fri, Nov 18, 2016 5:22 PM 42:10:00 5.00  
## 4 Chalfont Running Thu, Nov 17, 2016 6:01 AM 50:42:00 7.00  
## 5 Chalfont Running Wed, Nov 16, 2016 12:07 PM 38:49:00 5.01  
## 6 Chalfont Running Tue, Nov 15, 2016 5:55 AM 58:14:00 8.00  
## Elevation.Gain Avg.Speed.Avg.Pace. Max.Speed.Best.Pace. Avg.HR Max.HR  
## 1 529 7:08 3:04 151 200  
## 2 145 8:25 6:16 124 144  
## 3 238 8:26 7:07 133 148  
## 4 276 7:14 4:52 145 195  
## 5 140 7:45 6:21 135 164  
## 6 334 7:17 5:41 145 162  
## Calories sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF  
## 1 2,113 -- -- -- -- --  
## 2 329 -- -- -- -- --  
## 3 554 -- -- -- -- --  
## 4 762 -- -- -- -- --  
## 5 536 -- -- -- -- --  
## 6 853 -- -- -- -- --  
## Training.Effect  
## 1 --  
## 2 --  
## 3 --  
## 4 --  
## 5 --  
## 6 --

Okay, good, same amount of runs and variables. Now it's time to load in the rest of the Garmin data files in and combine them all into a single data frame to hold all of the Garmin data.

setwd("C:/Users/snewns/Dropbox/RunningAnalysis/Data")  
garmin2 <- read.csv("garmin2.csv")  
garmin3 <- read.csv("garmin3.csv")  
garmin4 <- read.csv("garmin4.csv")  
garmin5 <- read.csv("garmin5.csv")  
  
#Combine data files into 1 data frame  
garminFull <- Reduce(function(...) merge(..., all=TRUE), list(garmin1, garmin2, garmin3, garmin4,   
 garmin5, garmin6))  
dim(garminFull)

## [1] 122 16

summary(garminFull)

## Activity.Name Start   
## Chalfont Running :90 : 2   
## Running :19 Fri, Aug 5, 2016 5:30 AM : 1   
## Sea Isle City Running: 6 Fri, Jul 22, 2016 5:29 AM: 1   
## : 2 Fri, Jul 29, 2016 5:14 AM: 1   
## Pittsburgh Running : 2 Mon, Aug 8, 2016 7:31 AM : 1   
## Ocean City Running : 1 Sat, Aug 6, 2016 6:18 AM : 1   
## (Other) : 2 (Other) :115   
## Time Distance Elevation.Gain Avg.Speed.Avg.Pace.  
## : 2 Min. : 0.120 -- :16 7:45 : 5   
## 25:16:00: 2 1st Qu.: 5.010 0 : 6 7:43 : 5   
## 38:49:00: 2 Median :10.000 238 : 3 7:30 : 4   
## 1:02:54 : 1 Mean : 9.716 : 2 7:35 : 4   
## 1:04:59 : 1 3rd Qu.:13.000 224 : 2 7:48 : 4   
## 1:13:51 : 1 Max. :26.270 468 : 2 8:04 : 4   
## (Other) :113 NA's :2 (Other):91 (Other):96   
## Max.Speed.Best.Pace. Avg.HR Max.HR Calories   
## 4:52 : 4 Length:122 Length:122 : 2   
## 6:28 : 3 Class :character Class :character 1,090 : 2   
## 5:30 : 3 Mode :character Mode :character 1,085 : 2   
## 5:59 : 3 191 : 2   
## : 2 1,499 : 2   
## 4:57 : 2 1,018 : 1   
## (Other):105 (Other):111   
## sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF Training.Effect  
## : 2 : 2 : 2 : 2 : 2 : 2   
## --:120 --:120 --:120 --:120 --:120 --:120   
##   
##   
##   
##   
##

head(garminFull)

## Activity.Name Start Time Distance  
## 1 NA  
## 2 NA  
## 3 Chalfont Running Fri, Aug 5, 2016 5:30 AM 1:30:35 12.00  
## 4 Chalfont Running Fri, Jul 22, 2016 5:29 AM 1:24:31 11.00  
## 5 Chalfont Running Fri, Jul 29, 2016 5:14 AM 1:34:55 12.02  
## 6 Chalfont Running Sat, Jul 23, 2016 7:29 AM 48:25:00 6.00  
## Elevation.Gain Avg.Speed.Avg.Pace. Max.Speed.Best.Pace. Avg.HR Max.HR  
## 1   
## 2 <NA> <NA>  
## 3 459 7:33 5:05 148 158  
## 4 468 7:41 5:48 150 161  
## 5 477 7:54 5:35 149 160  
## 6 230 8:04 6:36 131 155  
## Calories sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF  
## 1   
## 2   
## 3 1,260 -- -- -- -- --  
## 4 1,170 -- -- -- -- --  
## 5 1,241 -- -- -- -- --  
## 6 721 -- -- -- -- --  
## Training.Effect  
## 1   
## 2   
## 3 --  
## 4 --  
## 5 --  
## 6 --

It looks like there's 2 random NULL records in the 122 activities, located at the start of the data frame's records. These are throwing things off, and maybe they're due the way Garmin delivers the data when I downloaded the activities from the Garmin Connect site. Or maybe *I* did something wrong. I don't know really.

But we still have to remove these top 2 blank rows

garminFull <- garminFull[-c(1,2),]  
head(garminFull)

## Activity.Name Start Time Distance  
## 3 Chalfont Running Fri, Aug 5, 2016 5:30 AM 1:30:35 12.00  
## 4 Chalfont Running Fri, Jul 22, 2016 5:29 AM 1:24:31 11.00  
## 5 Chalfont Running Fri, Jul 29, 2016 5:14 AM 1:34:55 12.02  
## 6 Chalfont Running Sat, Jul 23, 2016 7:29 AM 48:25:00 6.00  
## 7 Chalfont Running Sat, Jul 30, 2016 7:28 AM 49:29:00 6.01  
## 8 Chalfont Running Sun, Jul 24, 2016 8:17 AM 2:03:17 16.02  
## Elevation.Gain Avg.Speed.Avg.Pace. Max.Speed.Best.Pace. Avg.HR Max.HR  
## 3 459 7:33 5:05 148 158  
## 4 468 7:41 5:48 150 161  
## 5 477 7:54 5:35 149 160  
## 6 230 8:04 6:36 131 155  
## 7 244 8:14 7:01 134 150  
## 8 572 7:42 5:48 154 165  
## Calories sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF  
## 3 1,260 -- -- -- -- --  
## 4 1,170 -- -- -- -- --  
## 5 1,241 -- -- -- -- --  
## 6 721 -- -- -- -- --  
## 7 617 -- -- -- -- --  
## 8 1,653 -- -- -- -- --  
## Training.Effect  
## 3 --  
## 4 --  
## 5 --  
## 6 --  
## 7 --  
## 8 --

# Further Cleaning

## Remove Extra Columns

So now it's time to remove the Strava columns that would be unneccessary for what I'm doing here, based on the variables I pulled from Garmin.

keepCols <- NA  
keepCols <- c("Activity.Id", "When", "Type", "Gear", "Name", "Dist.mi", "Elv.ft", "Elapsed.Time", "Moving.Time",  
 "Speed.mph", "Pace..mi", "Max.Pace..mi", "Cad", "Heart", "Max.Heart", "Elev.Dist.ft.mi",  
 "Elev.Time.ft.h", "Cal", "Segs", "PRs", "Kudos")  
newStrava <- strava[keepCols]  
str(newStrava)

## 'data.frame': 201 obs. of 21 variables:  
## $ Activity.Id : int 780935487 780934297 779226144 777917114 777267101 776017804 775641778 774553806 773462012 772172603 ...  
## $ When : Factor w/ 201 levels "1/20/2016 6:18",..: 47 45 44 43 42 41 40 39 38 37 ...  
## $ Type : Factor w/ 1 level "Run": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Gear : Factor w/ 5 levels "ASICS dunno Black, Yellow, Red",..: 5 3 3 3 3 5 2 2 5 3 ...  
## $ Name : Factor w/ 19 levels "10k Tune Up Run",..: 15 16 16 4 16 4 16 11 4 19 ...  
## $ Dist.mi : num 43119 4832 8050 11270 8058 ...  
## $ Elv.ft : num 136 37.3 64.8 50.8 36.8 ...  
## $ Elapsed.Time : int 11443 1571 2616 3159 2404 3638 1046 5734 3215 3841 ...  
## $ Moving.Time : int 11234 1515 2530 3042 2328 3494 1046 5086 2994 3439 ...  
## $ Speed.mph : num 3.84 3.19 3.18 3.7 3.46 ...  
## $ Pace..mi : num 419 505 506 434 465 ...  
## $ Max.Pace..mi : num 159 424 383 293 358 ...  
## $ Cad : num 93.1 87.6 87.3 92.8 90.2 93.5 85.6 93.6 92.5 93.2 ...  
## $ Heart : num 151 124 133 145 135 ...  
## $ Max.Heart : int 200 144 148 194 164 162 118 157 154 167 ...  
## $ Elev.Dist.ft.mi: num 3.15 7.72 8.05 4.51 4.57 7.09 0 7.34 4.49 5.82 ...  
## $ Elev.Time.ft.h : num 43.6 88.6 92.2 60.1 56.9 ...  
## $ Cal : num 4227 474 789 1104 790 ...  
## $ Segs : int 17 0 0 0 0 0 0 4 0 9 ...  
## $ PRs : int 17 0 0 0 0 0 0 0 0 0 ...  
## $ Kudos : int 3 0 0 0 0 0 0 0 0 0 ...

Now I want to split the DateTime field **When** into a specific **Date** and **StartTime** fields.

library(stringr)  
newStrava$Date <- str\_split\_fixed(newStrava$When, " ", 2)[,1]  
newStrava$StartTime <- str\_split\_fixed(newStrava$When, " ", 2)[,2]  
  
str(newStrava)

## 'data.frame': 201 obs. of 23 variables:  
## $ Activity.Id : int 780935487 780934297 779226144 777917114 777267101 776017804 775641778 774553806 773462012 772172603 ...  
## $ When : Factor w/ 201 levels "1/20/2016 6:18",..: 47 45 44 43 42 41 40 39 38 37 ...  
## $ Type : Factor w/ 1 level "Run": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Gear : Factor w/ 5 levels "ASICS dunno Black, Yellow, Red",..: 5 3 3 3 3 5 2 2 5 3 ...  
## $ Name : Factor w/ 19 levels "10k Tune Up Run",..: 15 16 16 4 16 4 16 11 4 19 ...  
## $ Dist.mi : num 43119 4832 8050 11270 8058 ...  
## $ Elv.ft : num 136 37.3 64.8 50.8 36.8 ...  
## $ Elapsed.Time : int 11443 1571 2616 3159 2404 3638 1046 5734 3215 3841 ...  
## $ Moving.Time : int 11234 1515 2530 3042 2328 3494 1046 5086 2994 3439 ...  
## $ Speed.mph : num 3.84 3.19 3.18 3.7 3.46 ...  
## $ Pace..mi : num 419 505 506 434 465 ...  
## $ Max.Pace..mi : num 159 424 383 293 358 ...  
## $ Cad : num 93.1 87.6 87.3 92.8 90.2 93.5 85.6 93.6 92.5 93.2 ...  
## $ Heart : num 151 124 133 145 135 ...  
## $ Max.Heart : int 200 144 148 194 164 162 118 157 154 167 ...  
## $ Elev.Dist.ft.mi: num 3.15 7.72 8.05 4.51 4.57 7.09 0 7.34 4.49 5.82 ...  
## $ Elev.Time.ft.h : num 43.6 88.6 92.2 60.1 56.9 ...  
## $ Cal : num 4227 474 789 1104 790 ...  
## $ Segs : int 17 0 0 0 0 0 0 4 0 9 ...  
## $ PRs : int 17 0 0 0 0 0 0 0 0 0 ...  
## $ Kudos : int 3 0 0 0 0 0 0 0 0 0 ...  
## $ Date : chr "11/20/2016" "11/19/2016" "11/18/2016" "11/17/2016" ...  
## $ StartTime : chr "7:00" "16:48" "17:22" "6:01" ...

Okay, good, I've got two 2 character fields at the end there. Now onto the Garmin data cleaning. The date field, **Start** in Garmin is a bit different than in Strava, so some work needs to be done to get it into the same format by first creating new fields from **Start**.

library(tidyr)  
  
#Specify the new column names:  
vars <- c("Date", "StartTime")  
vars2 <- c("DOW", "Date")  
  
#Separate columns according to regex and/or delimiteres and proceed to drop remaining extra columns:  
garminFull <- separate(garminFull, Start, into = vars, sep = "(?<=6 )", extra = "merge", remove = TRUE)  
garminFull <- separate(garminFull, Date, into = vars2, sep = ", ", extra = "merge", remove = TRUE)  
  
#Create MonthNumber field based on what is in Date field  
garminFull$monthNum <- ifelse(grepl("Jan",garminFull$Date),1,  
 ifelse(grepl("Feb",garminFull$Date),2,  
 ifelse(grepl("Mar",garminFull$Date),3,  
 ifelse(grepl("Apr",garminFull$Date),4,  
 ifelse(grepl("May",garminFull$Date),5,  
 ifelse(grepl("Jun",garminFull$Date),6,  
 ifelse(grepl("Jul",garminFull$Date),7,  
 ifelse(grepl("Aug",garminFull$Date),8,  
 ifelse(grepl("Sep",garminFull$Date),9,  
 ifelse(grepl("Oct",garminFull$Date),10,  
 ifelse(grepl("Nov",garminFull$Date),11,  
 ifelse(grepl("Dec",garminFull$Date),12,NA))))))))))))  
head(garminFull,2)

## Activity.Name DOW Date StartTime Time Distance  
## 3 Chalfont Running Fri Aug 5, 2016 5:30 AM 1:30:35 12  
## 4 Chalfont Running Fri Jul 22, 2016 5:29 AM 1:24:31 11  
## Elevation.Gain Avg.Speed.Avg.Pace. Max.Speed.Best.Pace. Avg.HR Max.HR  
## 3 459 7:33 5:05 148 158  
## 4 468 7:41 5:48 150 161  
## Calories sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF  
## 3 1,260 -- -- -- -- --  
## 4 1,170 -- -- -- -- --  
## Training.Effect monthNum  
## 3 -- 8  
## 4 -- 7

Now I've got **Date, Start Time, Day of Week,** and **Month Number** columns, and I can seperate the **Date** column in **Day, Month,** and **Year**.

vars3 <- c("Month", "Date")  
vars4 <- c("Day", "Year")  
garminFull <- separate(garminFull, Date, into = vars3, sep = " ", extra = "merge", remove = TRUE)  
garminFull <- separate(garminFull, Date, into = vars4, sep = ", ", extra = "merge", remove = TRUE)  
head(garminFull,2)

## Activity.Name DOW Month Day Year StartTime Time Distance  
## 3 Chalfont Running Fri Aug 5 2016 5:30 AM 1:30:35 12  
## 4 Chalfont Running Fri Jul 22 2016 5:29 AM 1:24:31 11  
## Elevation.Gain Avg.Speed.Avg.Pace. Max.Speed.Best.Pace. Avg.HR Max.HR  
## 3 459 7:33 5:05 148 158  
## 4 468 7:41 5:48 150 161  
## Calories sumStrokes Avg.Strokes Min.Strokes Avg.SWOLF Best.SWOLF  
## 3 1,260 -- -- -- -- --  
## 4 1,170 -- -- -- -- --  
## Training.Effect monthNum  
## 3 -- 8  
## 4 -- 7

Looks like there's extra whitespace to the right in the **Year** field, so let's remove it and then create a Date field from its 3 components: Day, Month, and Year.

garminFull$Year <- trimws(garminFull$Year)  
  
#Create Date field from 3 components: Day, Month, Year  
garminFull$Date <- format(as.Date(with(garminFull, paste(Year, monthNum, Day,sep="-")), "%Y-%m-%d"), "%m/%d/%Y")  
str(garminFull$Date)

## chr [1:120] "08/05/2016" "07/22/2016" "07/29/2016" ...

Okay, so we've got a Date field, although it's a char data type. But I'll fix that later. For now, I'll remove my unneccesary Garmin columns.

keepColsGarmin <- NA  
keepColsGarmin <- c("DOW", "Month", "StartTime", "Time", "Distance", "Elevation.Gain", "Avg.Speed.Avg.Pace.", "Avg.HR",   
 "Max.HR", "Calories", "Date", "monthNum")  
newGarmin <- garminFull[keepColsGarmin]  
head(newGarmin,2)

## DOW Month StartTime Time Distance Elevation.Gain Avg.Speed.Avg.Pace.  
## 3 Fri Aug 5:30 AM 1:30:35 12 459 7:33  
## 4 Fri Jul 5:29 AM 1:24:31 11 468 7:41  
## Avg.HR Max.HR Calories Date monthNum  
## 3 148 158 1,260 08/05/2016 8  
## 4 150 161 1,170 07/22/2016 7

Now to change my format for my **Date** fields in both datasets, just as my personal preference.

newGarmin$Date <- as.Date(newGarmin$Date,"%m/%d/%Y")  
newStrava$Date <- as.Date(newStrava$Date,"%m/%d/%Y")  
  
#Check data type  
str(newGarmin$Date)

## Date[1:120], format: "2016-08-05" "2016-07-22" "2016-07-29" "2016-07-23" ...

str(newStrava$Date)

## Date[1:201], format: "2016-11-20" "2016-11-19" "2016-11-18" "2016-11-17" ...

Now that I've got these fields in the right format, I need to sort them so that when I combine the data sets, fields coming from both data sets are showing data for the *same* activity.

#Sort data frames from earliest to last date  
newGarmin <- newGarmin[order(newGarmin$Date, decreasing = FALSE),]   
newStrava <- newStrava[order(newStrava$Date, decreasing = FALSE),]  
head(newGarmin$Date)

## [1] "2016-07-19" "2016-07-20" "2016-07-21" "2016-07-22" "2016-07-23"  
## [6] "2016-07-24"

head(newStrava$Date)

## [1] "2016-01-20" "2016-01-30" "2016-01-31" "2016-02-02" "2016-02-03"  
## [6] "2016-02-04"

So, my runs from Strava seem to be starting back in January which was *way* before my marathon training plan kicked in, and was mainly for the Broad Street Run. The marathon training plan started right after the Tour De Shore bike ride I did on July 17, so I need to remove all runs prior to that 1st training plan run on July 19 (two days later since I needed one to recover from that ride!)

newStrava <- newStrava[!newStrava$Date < "2016-07-18",]

I need to double-check that the amount of rows in each data frame match up so that they can be combined correctly

nrow(newGarmin)

## [1] 120

nrow(newStrava)

## [1] 117

So there's still 3 extra Garmin runs, so I need to find out what they are. I'll check a table of the distance to see if anything sticks out.

table(newGarmin$Distance)

##   
## 0.12 0.16 1 1.42 2 2.06 3 4 4.01 5 5.01 6   
## 1 1 2 1 1 1 5 5 3 5 6 4   
## 6.01 6.51 6.52 7 8 8.01 9.01 9.33 10 10.01 11 11.01   
## 6 1 1 5 2 2 3 1 8 6 7 3   
## 12 12.01 12.02 13 13.01 14 14.01 15 15.01 16.02 17.01 17.02   
## 4 3 1 3 2 2 3 3 6 1 1 1   
## 18 20 20.01 21 22 24 26.27   
## 1 3 1 1 2 1 1

So there's 2 runs that are less than one-fifth of a mile? That is *definitely* not right. The run with **1.42** miles seems but off as well maybe.

newGarmin[newGarmin$Distance=="0.12",]

## DOW Month StartTime Time Distance Elevation.Gain  
## 98 Sat Aug 12:06 PM 59:00:00 0.12 --  
## Avg.Speed.Avg.Pace. Avg.HR Max.HR Calories Date monthNum  
## 98 --:-- 109 123 389 2016-08-27 8

newGarmin[newGarmin$Distance=="0.16",]

## DOW Month StartTime Time Distance Elevation.Gain  
## 101 Thu Sep 5:50 AM 36:23:00 0.16 --  
## Avg.Speed.Avg.Pace. Avg.HR Max.HR Calories Date monthNum  
## 101 --:-- 101 114 227 2016-09-01 9

newGarmin[newGarmin$Distance=="1.42",]

## DOW Month StartTime Time Distance Elevation.Gain Avg.Speed.Avg.Pace.  
## 116 Fri Aug 6:12 PM 11:33 1.42 -- 8:07  
## Avg.HR Max.HR Calories Date monthNum  
## 116 -- -- 191 2016-08-12 8

So those 1st two runs look to actually be spin sessions, based on the long times but incredibly short distances (thanks to my wristwatch moving up and down while I was tracking heart rate on the bike), but the 1.42 mile run just seems to be a light run that I didn't run to an even 0.50 interval of mile (*very* unlike me). So let's remove those 1st two runs, but keep the 1.42-miler.

newGarmin <- newGarmin[!newGarmin$Date == "2016-08-27",]  
newGarmin <- newGarmin[!newGarmin$Date == "2016-09-01",]

But there's still 1 more extra run. Let's check the dates for any duplicates.

table(newGarmin$Date) > 1

##   
## 2016-07-19 2016-07-20 2016-07-21 2016-07-22 2016-07-23 2016-07-24   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-07-26 2016-07-27 2016-07-28 2016-07-29 2016-07-30 2016-07-31   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-08-02 2016-08-03 2016-08-04 2016-08-05 2016-08-06 2016-08-07   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-08-08 2016-08-10 2016-08-11 2016-08-12 2016-08-15 2016-08-16   
## FALSE FALSE FALSE TRUE FALSE FALSE   
## 2016-08-17 2016-08-18 2016-08-19 2016-08-20 2016-08-21 2016-08-22   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-08-23 2016-08-24 2016-08-25 2016-08-26 2016-08-28 2016-08-29   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-08-30 2016-08-31 2016-09-02 2016-09-03 2016-09-04 2016-09-06   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-09-07 2016-09-08 2016-09-09 2016-09-10 2016-09-11 2016-09-12   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-09-13 2016-09-14 2016-09-15 2016-09-16 2016-09-17 2016-09-18   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-09-19 2016-09-20 2016-09-21 2016-09-22 2016-09-23 2016-09-25   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-09-26 2016-09-27 2016-09-28 2016-09-29 2016-09-30 2016-10-01   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-10-02 2016-10-03 2016-10-04 2016-10-05 2016-10-06 2016-10-07   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-10-08 2016-10-09 2016-10-10 2016-10-11 2016-10-12 2016-10-13   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-10-14 2016-10-15 2016-10-16 2016-10-17 2016-10-18 2016-10-19   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-10-20 2016-10-21 2016-10-22 2016-10-23 2016-10-24 2016-10-25   
## FALSE FALSE FALSE TRUE FALSE FALSE   
## 2016-10-26 2016-10-27 2016-10-28 2016-10-29 2016-10-30 2016-10-31   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-11-01 2016-11-02 2016-11-03 2016-11-04 2016-11-05 2016-11-06   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-11-07 2016-11-08 2016-11-09 2016-11-10 2016-11-11 2016-11-12   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-11-13 2016-11-14 2016-11-15 2016-11-16 2016-11-17 2016-11-18   
## FALSE FALSE FALSE FALSE FALSE FALSE   
## 2016-11-19 2016-11-20   
## FALSE FALSE

There was probably an easier way to do this, but I see that there's the 2 runs on August 12, and also 2 on October 23.

which(newStrava$Date == '8/12/2016') #runs 95 and 96

## integer(0)

Let's check those records.

newStrava[95:96,]

## Activity.Id When Type Gear Name  
## 23 759408079 10/29/2016 10:09 Run Brooks Ghost Red/Black GA Run  
## 22 760801450 10/30/2016 9:13 Run Hoka One One Clifton Reds Long Run  
## Dist.mi Elv.ft Elapsed.Time Moving.Time Speed.mph Pace..mi Max.Pace..mi  
## 23 11269.6 48.0 3244 3144 3.5845 449.0070 328.4349  
## 22 35408.8 205.4 11757 9772 3.6235 444.1789 335.2746  
## Cad Heart Max.Heart Elev.Dist.ft.mi Elev.Time.ft.h Cal Segs PRs  
## 23 92.7 144.3 161 4.26 55.0 1103.4 0 0  
## 22 84.8 139.9 155 5.80 75.7 3471.1 4 0  
## Kudos Date StartTime  
## 23 0 2016-10-29 10:09  
## 22 0 2016-10-30 9:13

Looks like I just ran twice that day. I know I was down the shore, maybe I was just bored? Now to check the October 23 runs.

newGarmin[newGarmin$Date == "2016-10-23",]

## DOW Month StartTime Time Distance Elevation.Gain  
## 69 Sun Oct 11:43 AM 1:04:28 15.01 657  
## 108 Sun Oct 1:08 PM 7:53 1.00 --  
## Avg.Speed.Avg.Pace. Avg.HR Max.HR Calories Date monthNum  
## 69 4:18 114 146 510 2016-10-23 10  
## 108 7:52 119 128 95 2016-10-23 10

While I'd love to run a 4:18 min/mile pace and finish 15 miles in an hour, that looks like another bike ride, but outdoors , since it has a value for elevation, unlike our spin rides above. So let's remove that one and then check the row counts again.

newGarmin <- newGarmin[!(newGarmin$Date == "2016-10-23" & newGarmin$Elevation.Gain == 657),]  
  
nrow(newGarmin)

## [1] 117

nrow(newStrava)

## [1] 117

Great, now I have the same amount of runs in each data set. Let's check the first couple of records just to make sure they line up.

head(newGarmin)

## DOW Month StartTime Time Distance Elevation.Gain  
## 12 Tue Jul 5:51 AM 1:04:59 9.01 351  
## 15 Wed Jul 5:36 AM 1:23:29 11.00 461  
## 94 Thu Jul 6:45 AM 32:40:00 4.00 --  
## 4 Fri Jul 5:29 AM 1:24:31 11.00 468  
## 6 Sat Jul 7:29 AM 48:25:00 6.00 230  
## 8 Sun Jul 8:17 AM 2:03:17 16.02 572  
## Avg.Speed.Avg.Pace. Avg.HR Max.HR Calories Date monthNum  
## 12 7:13 149 167 943 2016-07-19 7  
## 15 7:35 150 160 1,176 2016-07-20 7  
## 94 8:10 139 149 442 2016-07-21 7  
## 4 7:41 150 161 1,170 2016-07-22 7  
## 6 8:04 131 155 721 2016-07-23 7  
## 8 7:42 154 165 1,653 2016-07-24 7

head(newStrava)

## Activity.Id When Type Gear  
## 117 646029304 7/19/2016 5:51 Run Brooks Ghost Red/Black  
## 116 647216014 7/20/2016 5:36 Run Brooks Ghost Red/Black  
## 115 648410753 7/21/2016 6:45 Run ASICS dunno Black, Yellow, Red  
## 114 649445500 7/22/2016 5:29 Run Brooks Ghost Red/Black  
## 113 650512799 7/23/2016 7:29 Run ASICS dunno Black, Yellow, Red  
## 112 651933678 7/24/2016 8:17 Run Brooks Ghost Red/Black  
## Name Dist.mi Elv.ft Elapsed.Time Moving.Time Speed.mph  
## 117 Classic LT Run 14496.0 90.2 4415 3899 3.7179  
## 116 MIddle Long Run 17706.6 112.9 5180 5009 3.5350  
## 115 Recovery Run 6440.2 0.0 1959 1959 3.2875  
## 114 GA Run 17710.0 113.5 5642 5070 3.4931  
## 113 Morning Run 9656.8 45.6 3101 2904 3.3253  
## 112 Long Run 25773.8 147.3 8358 7397 3.4844  
## Pace..mi Max.Pace..mi Cad Heart Max.Heart Elev.Dist.ft.mi  
## 117 432.9135 303.6510 84.2 148.9 166 6.22  
## 116 455.2834 357.6284 84.2 149.7 160 6.38  
## 115 489.5624 459.8057 85.6 139.0 149 0.00  
## 114 460.7552 315.5602 88.0 150.3 161 6.41  
## 113 483.9297 342.4201 86.9 130.7 155 4.72  
## 112 461.8817 328.4349 85.0 154.1 164 5.72  
## Elev.Time.ft.h Cal Segs PRs Kudos Date StartTime  
## 117 83.3 1424.0 2 0 0 2016-07-19 5:51  
## 116 81.1 1738.4 4 0 0 2016-07-20 5:36  
## 115 0.0 628.1 0 0 0 2016-07-21 6:45  
## 114 80.6 1739.8 4 0 0 2016-07-22 5:29  
## 113 56.5 947.3 0 0 0 2016-07-23 7:29  
## 112 71.7 2531.4 5 0 0 2016-07-24 8:17

Awesome, they line up. Let's rename the date and time columns in Garmin so we can distinguish them from the respective Strava fields and then remove the unneccessary **When** column

library(plyr)  
newGarmin <- rename(newGarmin, c("Date"="date\_garmin"))  
newGarmin <- rename(newGarmin, c("StartTime"="StartTime\_AM\_PM"))  
newStrava$When <- NULL #removes column

Now that we have cleaned up both data sets, and can combine them into one, full, complete data set, and double-check the dates again.

newFullData <- cbind(newGarmin,newStrava)  
#inspect  
head(newFullData,3)

## DOW Month StartTime\_AM\_PM Time Distance Elevation.Gain  
## 12 Tue Jul 5:51 AM 1:04:59 9.01 351  
## 15 Wed Jul 5:36 AM 1:23:29 11.00 461  
## 94 Thu Jul 6:45 AM 32:40:00 4.00 --  
## Avg.Speed.Avg.Pace. Avg.HR Max.HR Calories date\_garmin monthNum  
## 12 7:13 149 167 943 2016-07-19 7  
## 15 7:35 150 160 1,176 2016-07-20 7  
## 94 8:10 139 149 442 2016-07-21 7  
## Activity.Id Type Gear Name Dist.mi  
## 12 646029304 Run Brooks Ghost Red/Black Classic LT Run 14496.0  
## 15 647216014 Run Brooks Ghost Red/Black MIddle Long Run 17706.6  
## 94 648410753 Run ASICS dunno Black, Yellow, Red Recovery Run 6440.2  
## Elv.ft Elapsed.Time Moving.Time Speed.mph Pace..mi Max.Pace..mi Cad  
## 12 90.2 4415 3899 3.7179 432.9135 303.6510 84.2  
## 15 112.9 5180 5009 3.5350 455.2834 357.6284 84.2  
## 94 0.0 1959 1959 3.2875 489.5624 459.8057 85.6  
## Heart Max.Heart Elev.Dist.ft.mi Elev.Time.ft.h Cal Segs PRs Kudos  
## 12 148.9 166 6.22 83.3 1424.0 2 0 0  
## 15 149.7 160 6.38 81.1 1738.4 4 0 0  
## 94 139.0 149 0.00 0.0 628.1 0 0 0  
## Date StartTime  
## 12 2016-07-19 5:51  
## 15 2016-07-20 5:36  
## 94 2016-07-21 6:45

#check dates  
head(newFullData[,c("Date","date\_garmin")])

## Date date\_garmin  
## 12 2016-07-19 2016-07-19  
## 15 2016-07-20 2016-07-20  
## 94 2016-07-21 2016-07-21  
## 4 2016-07-22 2016-07-22  
## 6 2016-07-23 2016-07-23  
## 8 2016-07-24 2016-07-24

tail(newFullData[,c("Date","date\_garmin")])

## Date date\_garmin  
## 88 2016-11-15 2016-11-15  
## 90 2016-11-16 2016-11-16  
## 85 2016-11-17 2016-11-17  
## 78 2016-11-18 2016-11-18  
## 81 2016-11-19 2016-11-19  
## 122 2016-11-20 2016-11-20

Okay, now I can move to keep the columns that I need for what I want to do with this data, and then order them as I like.

newFullData <- rename(newFullData, c("Activity.Id"="ID"))  
  
keepColsFull <- NA  
keepColsFull <- c("ID", "Gear", "Name", "Speed.mph", "Cad", "Date", "StartTime", "DOW", "Month", "Time", "Distance",   
 "Elevation.Gain", "Avg.Speed.Avg.Pace.", "Avg.HR", "Max.HR", "Calories", "monthNum")  
newFullData <- newFullData[keepColsFull]  
  
#rearrange columns  
newFullData <- newFullData[, c("ID", "Name", "Gear", "Date", "Month", "monthNum", "DOW", "StartTime", "Distance",   
 "Time", "Avg.Speed.Avg.Pace.", "Speed.mph", "Cad", "Elevation.Gain", "Avg.HR", "Max.HR",   
 "Calories")]

Now I want to check the names of the runs I input into Strava when uploading them from my watch to make sure they're consistent, and I then can categorize them for an accurate analysis.

table(newFullData$Name)

##   
## 10k Tune Up Run   
## 1   
## Afternoon Run   
## 0   
## Classic LT Run   
## 1   
## GA Run   
## 24   
## Long Run   
## 11   
## LT Intervals Run   
## 1   
## LT Run   
## 1   
## Lunch Run   
## 0   
## Middle Long Run   
## 16   
## MIddle Long Run   
## 2   
## ML Run   
## 3   
## Morning Long Run   
## 0   
## Morning Run   
## 2   
## MP Long Run   
## 2   
## Oh My God So Freakin Windy Philly Marathon   
## 1   
## Recovery Run   
## 40   
## Tempo Run   
## 5   
## Tune-Up Run   
## 1   
## VO2 Run   
## 6

So there seems to be some inconsistencies (and a typo) for my *Middle Long Runs*, so now I need to correct them. I'll put them all al **ML Run**.

newFullData$Name[newFullData$Name == 'MIddle Long Run'] <- 'ML Run'  
newFullData$Name[newFullData$Name == 'Middle Long Run'] <- 'ML Run'

There also seems to be 2 runs with the name "Morning Run", which was not the name of 1 of my types of runs, so let's investigate these runs.

newFullData[newFullData$Name == 'Morning Run',]

## ID Name Gear Date Month  
## 6 650512799 Morning Run ASICS dunno Black, Yellow, Red 2016-07-23 Jul  
## 16 655096239 Morning Run Brooks Ghost Red/Black 2016-07-27 Jul  
## monthNum DOW StartTime Distance Time Avg.Speed.Avg.Pace. Speed.mph  
## 6 7 Sat 7:29 6 48:25:00 8:04 3.3253  
## 16 7 Wed 5:15 13 1:40:41 7:45 3.4642  
## Cad Elevation.Gain Avg.HR Max.HR Calories  
## 6 86.9 230 131 155 721  
## 16 87.4 485 146 158 1,318

So, based off of distance, shoe type, and pace, I can safely assume that the run on 7/23 was a Recovery Run, and the run on 7/27 was a ML Run, so I can edit them as so.

newFullData$Name[newFullData$ID == 650512799] <- 'Recovery Run'  
newFullData$Name[newFullData$ID == 655096239] <- 'ML Run'

Now, based off of the run names, I want to create 4 categories (same as for Strava) for **workouts, Long Runs, Recovery Runs,** and just plain ol' **Runs**, which consist of **General Aerobic (GA)** and **Middle Long (ML) runs**

newFullData$RunType <- ifelse(grepl('LT',newFullData$Name),'Workout',  
 ifelse(grepl('Tempo',newFullData$Name),'Workout',  
 ifelse(grepl('Tune',newFullData$Name),'Workout',  
 ifelse(grepl('VO2',newFullData$Name),'Workout',  
 #ifelse(grepl('MP',newFullData$Name),'Workout',  
 ifelse(grepl('Long',newFullData$Name),'Long Run',  
 ifelse(grepl('Recovery',newFullData$Name),'Recovery Run',  
 ifelse(grepl('Marathon',newFullData$Name),'Race','Run')))))))#)  
table(newFullData$RunType)

##   
## Long Run Race Recovery Run Run Workout   
## 13 1 41 46 16

Looks like I've got a lot of "plain" runs and a lot of recovery runs. Just goes to show that you don't need to kill yourself out there every time you run to race a marathon! Now to save this dataset to a file to read into another script so I can do some analysis.

write.csv(newFullData, file = "cleanedMarathonTrainingData.csv", row.names = TRUE)