My Running Data

Dan Spencer

Background

I got into running a few years ago after spending too many long days working at a desk. In December 2016, I began tracking my runs using a running app called Strava. I started using this particular run-tracking app because I found that it does a pretty good job keeping track of all sorts of data, which you can then access yourself at a later date. In this particular document, I'll be cleaning my data aggregated by workout, creating a visualization to see if my speed is changing significantly over time, and then comparing my different runs to determine whether different shoes have a significant effect on my pace.

Cleaning the Data

I downloaded my data directly from Strava, and it came in a folder called `export_18831425`. Within this folder area number of spreadsheets and subfolders. I will focus on the `activities.csv` spreasheet for my analysis here.

```
activities <- read.csv("export_18831425/activities.csv")
names(activities) # This gives the names of the variables in the spreadsheet
```

```
##  [1] "Activity.ID"          "Activity.Date"        "Activity.Name"
##  [4] "Activity.Type"        "Activity.Description" "Elapsed.Time"
##  [7] "Distance"             "Relative.Effort"      "Commute"
## [10] "Activity.Gear"        "Filename"
```

```
table(activities$Activity.Type)
```

```
##
##  E-Bike Ride         Ride          Run Virtual Ride
##            3            6          366            1
```

My efforts today will focus on my running activities, and I'll only need to look at the `Activity.Date`, `Activity.Gear`, `Elapsed.Time`, and `Distance` variables, so I will begin by cleaning the data to match my needs.

```
library(tidyverse)
run_df <- activities %>%
  dplyr::filter(Activity.Type == "Run") %>% # Only look at runs
  dplyr::select(Activity.Date, Activity.Gear, Elapsed.Time, Distance)
head(run_df)
```

```
##               Activity.Date Activity.Gear Elapsed.Time Distance
## 1  Dec 14, 2016, 5:22:14 PM                       2208     8.05
## 2  Dec 16, 2016, 4:48:25 PM                       2230     7.95
## 3 Dec 17, 2016, 10:20:49 PM                       2102     7.94
## 4  Dec 18, 2016, 5:43:09 PM                       2169     8.04
## 5  Dec 21, 2016, 2:41:48 PM                       2750     9.43
## 6  Dec 22, 2016, 4:36:18 PM                       3063    10.44
```

Something that I notice here is that the `Activity.Date` variable is not in a standard format. Since I'm only interested in the date when the runs took place, I can try to format the dates into a date format that can be recognized by R. I can also convert the `Elapsed.Time` from seconds to minutes, convert `Distance` from

kilometers to miles, and find my pace by dividing the time in minutes by the distance in miles. I'll sum up the distances and times for any activities that occur on the same day to ensure I have a unique daily value.

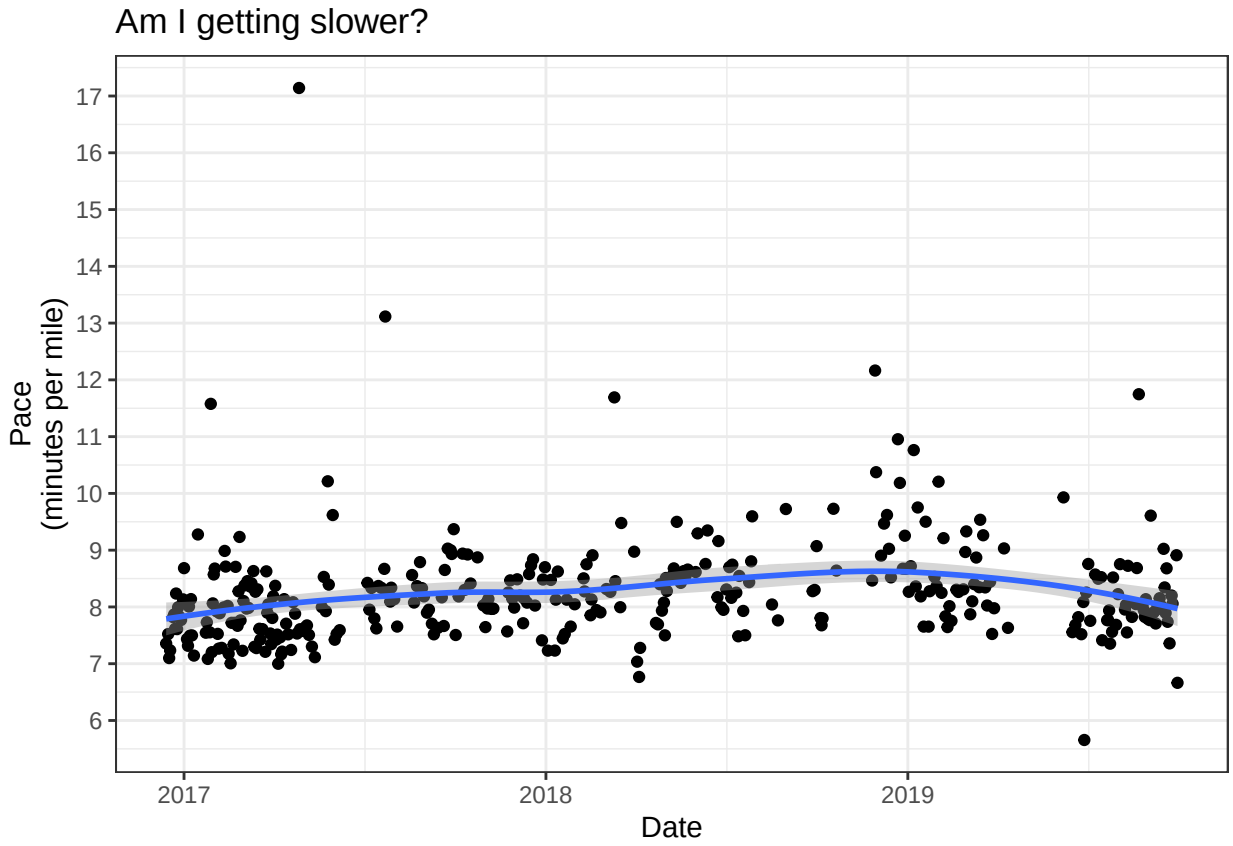
```
final_run_df <- run_df %>%
  dplyr::mutate(Date = substr(Activity.Date,1,12), # Get rid of the time information
                Date = as.Date(Date,format = "%b %d, %Y")) %>%  # Get R to read as a date
  dplyr::group_by(Date) %>% # Next, in case there are multiple runs in a day, group all of the data tog
  dplyr::summarize(Elapsed.Time = sum(Elapsed.Time),
                   Distance = sum(Distance),
                   Activity.Gear = first(Activity.Gear)) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(Distance = Distance / 1.609, # Convert from km to miles
                Minutes = Elapsed.Time / 60, # Convert from second to minutes
                Pace = Minutes / Distance)
```

Visualizing My Pace Through Time

It might be interesting to see if I have gotten faster or slower through time over the past few years. Am I getting slower as I get older?

```
ggplot(final_run_df) +
  geom_point(aes(x = Date, y = Pace)) + # Put the points on the graph
  geom_smooth(aes(x = Date, y = Pace)) + # Plot a smoothed curve over the points
  scale_y_continuous("Pace\n(minutes per mile)",breaks = seq(6,17)) + # Control where the grid lines ar
  ggtitle("Am I getting slower?") + # Adds a title
  theme_bw() # Gives a nice, clean plot with minimal color
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



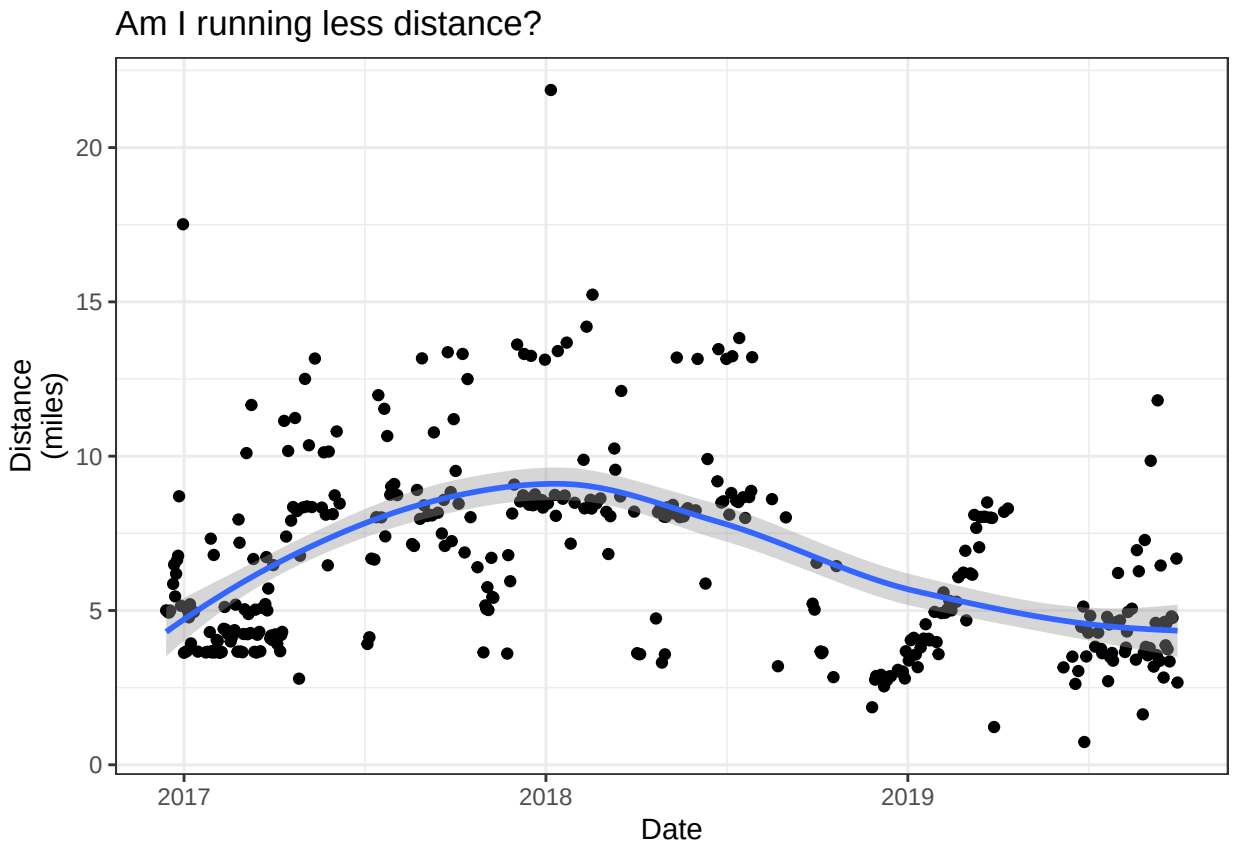
These data suggest that I was getting slower through time until around the beginning of 2019, when the trend reversed slightly. There are some hidden pieces of information here. First, I suffered an injury to my ankle in the summer of 2018, which slowed me down significantly as I recovered. I also suffered a mild muscle strain around March of 2019, which stopped my running completely for about two months. However, after my recovery I have taken smaller runs at a faster pace.

Visualizing Distances per Day Through Time

Something that has a large effect on pace is distance traveled. To this end, I can examine how the distances I have run have changed over time.

```
ggplot(final_run_df) +
  geom_point(aes(x = Date, y = Distance)) + # Put the points on the graph
  geom_smooth(aes(x = Date, y = Distance)) + # Plot a smoothed curve over the points
  scale_y_continuous("Distance\n(miles)") + # Control where the grid lines are on the y-axis
  ggtitle("Am I running less distance?") + # Adds a title
  theme_bw() # Gives a nice, clean plot with minimal color
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From this plot, it's clear that I am running less on average in a given day. My injuries and recoveries are also clearly observable, as my distance drops down after each injury.

Do the Shoes Make a Difference?

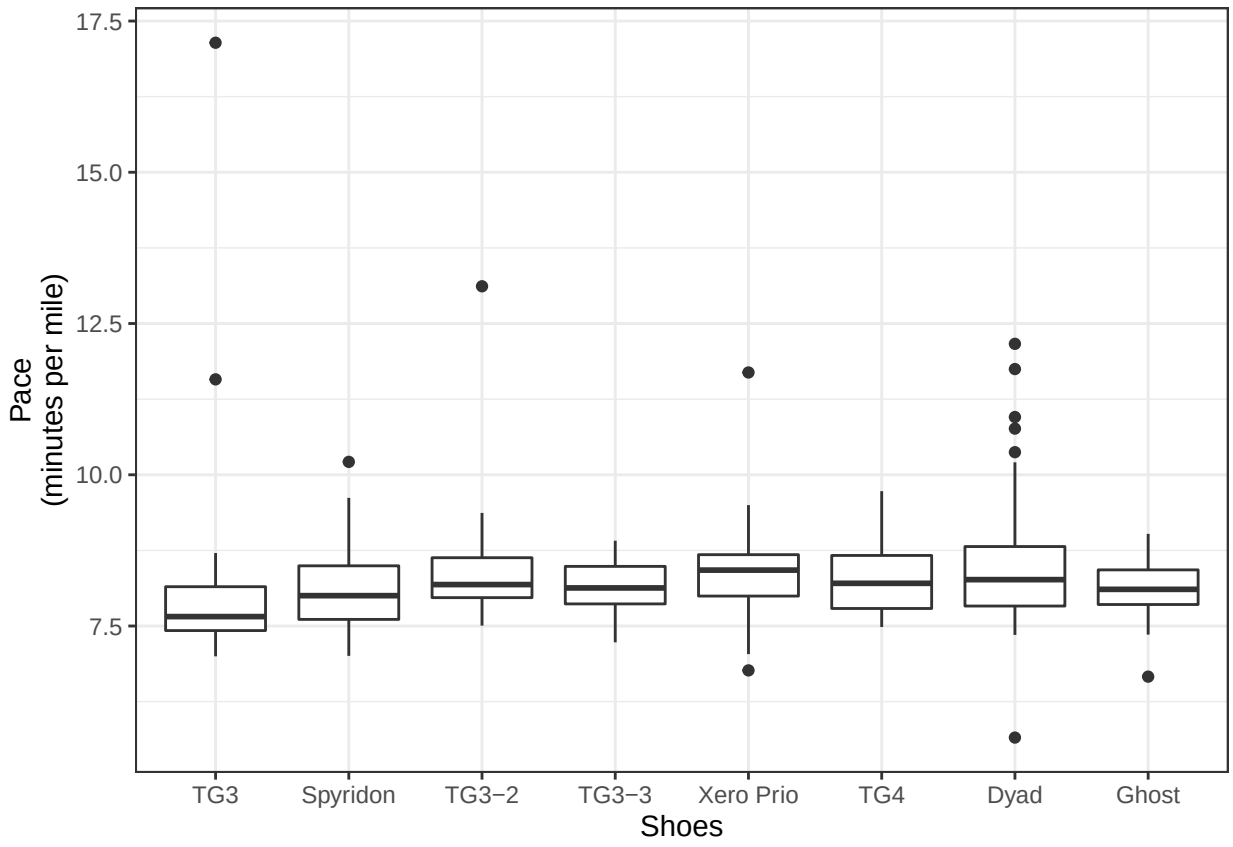
Now I'll see if I given pair of shoes makes me faster or slower. I'll begin by looking at a visualization of the data to see if there could be a difference.

```
shoes_df <- final_run_df %>%
  dplyr::filter(Activity.Gear != "") %>% # Get rid of runs with no shoes listed
  # Next, I'll change a shoe name to the correct name
  dplyr::mutate(Activity.Gear = ifelse(Activity.Gear == "12", "Ghost", as.character(Activity.Gear)))

# This part is important, as listing the shoes used by date may
# illustrate some patterns
date_order_shoes <- shoes_df %>%
  dplyr::group_by(Activity.Gear) %>%
  dplyr::summarize(first_worn = first(Date)) %>% # Find the first date a given shoe was worn
  dplyr::ungroup() %>%
  dplyr::mutate(date_rank = rank(first_worn)) %>% # Find the date rank for when shoes were first worn
  dplyr::arrange(date_rank) # Arrange the data frame by the date rank

shoes_df %>%
  dplyr::mutate(Activity.Gear = factor(as.character(Activity.Gear), # Changes the factor to order by da
                                       levels = as.character(date_order_shoes$Activity.Gear))) %>%
  ggplot() +
  geom_boxplot(aes(y = Pace, x = Activity.Gear)) +
```

```
  labs(y = "Pace\n(minutes per mile)", x = "Shoes") +
  theme_bw()
```



There might be a difference here. To see if there is a statistically-significant difference, I'll run a basic analysis of variance (ANOVA).

```
anova_shoes <- aov(Pace ~ Activity.Gear, data = shoes_df)
summary(anova_shoes)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Activity.Gear   7  14.71  2.1011   2.367 0.0226 *
## Residuals     334 296.45  0.8876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This shows that there is some significant difference in at least two different pairs of shoes, however, this does not tell use which shoes have significantly different average paces. In order to find that information, we will need to use something like Tukey's Honestly Significant Difference Test.

```
tukey_test <- TukeyHSD(anova_shoes)
tukey_test
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Pace ~ Activity.Gear, data = shoes_df)
##
## $Activity.Gear
##                              diff         lwr         upr     p adj
```

```
## Ghost-Dyad         -0.4011169249 -1.28368211  0.48144826 0.8631030
## Spyridon-Dyad      -0.3595597242 -0.91459663  0.19547718 0.4997251
## TG3-Dyad           -0.5366822583 -0.99732359 -0.07604092 0.0101903
## TG3-2-Dyad         -0.1196867894 -0.65575490  0.41638133 0.9974443
## TG3-3-Dyad         -0.3601830504 -0.93779258  0.21742648 0.5504586
## TG4-Dyad           -0.1499766619 -0.85966112  0.55970779 0.9981989
## Xero Prio-Dyad     -0.0423088266 -0.60261180  0.51799414 0.9999983
## Spyridon-Ghost      0.0415572006 -0.91001573  0.99313013 1.0000000
## TG3-Ghost          -0.1355653334 -1.03535171  0.76422104 0.9998043
## TG3-2-Ghost         0.2814301355 -0.65920481  1.22206508 0.9847405
## TG3-3-Ghost         0.0409338745 -0.92397951  1.00584726 1.0000000
## TG4-Ghost           0.2511402630 -0.79818195  1.30046247 0.9960444
## Xero Prio-Ghost     0.3588080983 -0.59584602  1.31346222 0.9459103
## TG3-Spyridon       -0.1771225340 -0.75915361  0.40490854 0.9831562
## TG3-2-Spyridon      0.2398729349 -0.40350779  0.88325366 0.9481472
## TG3-3-Spyridon     -0.0006233261 -0.67900560  0.67775894 1.0000000
## TG4-Spyridon        0.2095830624 -0.58428202  1.00344814 0.9927616
## Xero Prio-Spyridon  0.3172508977 -0.34645774  0.98095953 0.8290033
## TG3-2-TG3           0.4169954689 -0.14697549  0.98096642 0.3217738
## TG3-3-TG3           0.1764992079 -0.42709584  0.78009426 0.9866574
## TG4-TG3             0.3867055964 -0.34428431  1.11769550 0.7418769
## Xero Prio-TG3       0.4943734317 -0.09268162  1.08142848 0.1712335
## TG3-3-TG3-2        -0.2404962610 -0.90344842  0.42245590 0.9551390
## TG4-TG3-2          -0.0302898725 -0.81101058  0.75043084 1.0000000
## Xero Prio-TG3-2     0.0773779628 -0.57055121  0.72530714 0.9999593
## TG4-TG3-3           0.2102063885 -0.59960135  1.02001412 0.9934705
## Xero Prio-TG3-3     0.3178742238 -0.36482334  1.00057178 0.8474831
## Xero Prio-TG4       0.1076678353 -0.68988794  0.90522361 0.9999065
```

This shows that only two shoes have average paces that are significantly different: the TG3 (Merrill Trail Glove 3), and the Brooks Dyad. This is likely because I purchased the Dyad in order to recover from my planar fasciitis in late 2018.

Conclusion

I found that, while my pace was getting slower through the beginning of 2019, this is probably because my runs were getting longer and/or I was recovering from an injury. Otherwise, there does not seem to be a significant linear trend through time in terms of my average pace. From this, the take-home lesson seems to be, "Try to avoid an injury."