Beat Yourself: New Haven Road Race 5K Times Compared to Self

Michael Li and Davey Proctor

I. Introduction

The overall goal of this report is to explore/analyze the past results of the Faxon Law New Haven Road Race 5K data and hopefully uncover interesting trends and relationships. In total, we have 7 years (2012-2018) worth of data on the runners. Each year we have most of the runners' name, division, net time, age, and sex - in total this came out to almost eleven thousand observations (after cleaning). Of course the most important variable that was at the heart of our exploration was each runner's net time as it was the only objective measure of performance that we had.

We focused on evaluating each individual runner's performance over time (relative to himself or herself). Contrary to an expectation of self-improvement; runners worsen over time in general. This is perhaps best explained by the ravages of age. Other explanations include that the 5K is a pool of majority casual runners – and those casual runners contribute most to slip in net time year by year.

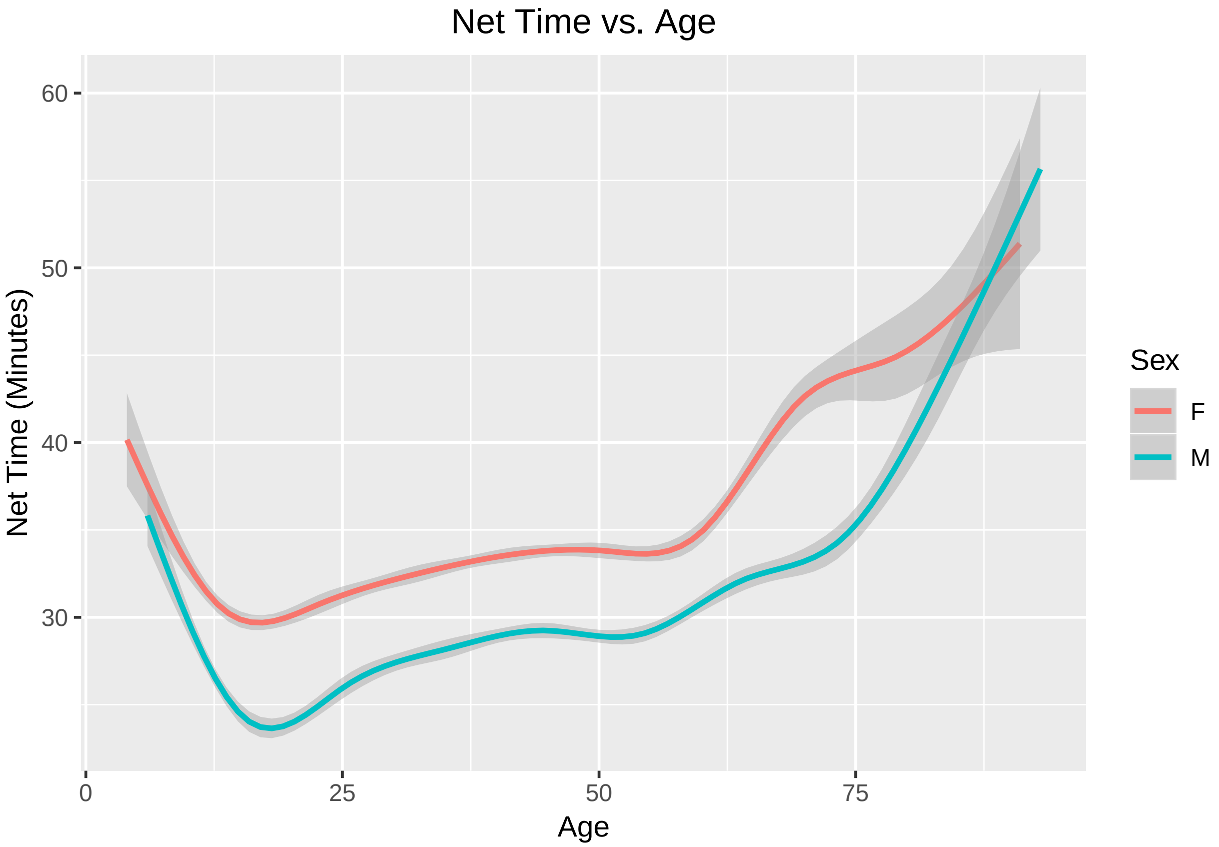
II. Data

A. Cleaning

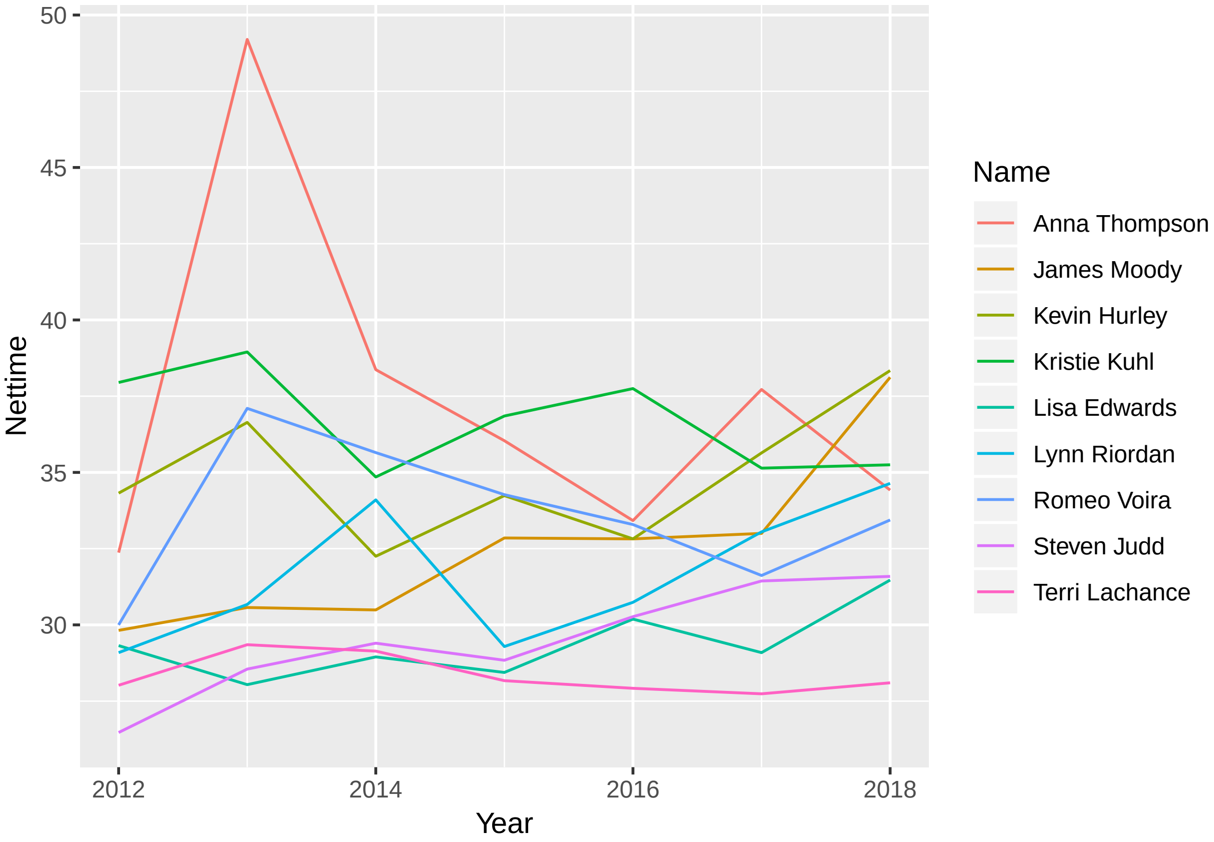
The cleaning process that we used for this dataset was actually quite involved. After scraping the data from the New Haven Road Race Website, we eliminated white space, standardized the format of the data across all 7 years, converted "Nettime" from an "hh:mm" format to minutes, and then combined all 7 years of data into one data frame. This initial cleaning step put our data into a usable format and also got rid of clearly erroneous entries.

From there, we were still left with many missing values. Luckily we were able to impute/infer many of the missing values from either previous years' data for each person or values from other related columns. We inferred age for many runners by using an age value they had in other years and added (or subtracted) the appropriate number of years from that existing age value. For the runners with missing "Sex" data, we were able to infer what their sex was from the first letter of the division column. Dealing with the issue of multiple runners with the same exact first and last name was a little more complicated. Since there was no unique ID for runners, we decided to eliminate all runners with the same name running within the same year.

B. Exploration



Since "Nettime" is the variable that lies at the center of our analysis, we start off by taking a very simple slice of the data and looking at net time vs. Age. This is displaying the conditional mean net time of men or women at each age. The grey band around the lines are the bands for the 95% confidence interval. What is interesting to see here is that it seems like there are generally 3 broad sections we could divide this graph into. Runners less than 20 years old seem to improve their net times significantly, with each year with 18/19 year olds on average being the fastest of all runners. From 25 to 50 years old, the average net time slowly increases year to year. From 50 onward, the average net time increases a lot each year. However, these net times are averaged over all runners from 2012 to 2018 - this means that this graph does NOT show the relative improvement of individuals over time. For example, it is very plausible that an avid runner who runs in this race each year could actually improve his or her net time from age 50 onward, however because the average 25 year old is faster than the average 50 year old, this graph would not show that trend.



Just to get a sense of what the data for "relativized" runners would look like, this graph shows the performance over time of a sample of 10 individual "super runners" (runners who ran the race all 7 years from 2012 to 2018).

Moving forward, our analysis was based around exploring this idea of "relativized" runner performances. On the surface, it seemed intuitive that runners who ran this race for multiple years would see improvements in their net times. To accomplish this we fit a linear model to each runners' net times over the years. The slope, intercept, and standard errors of the linear models for each runners provided us a way of summarizing the general (average) trend of each runner's net time over time.

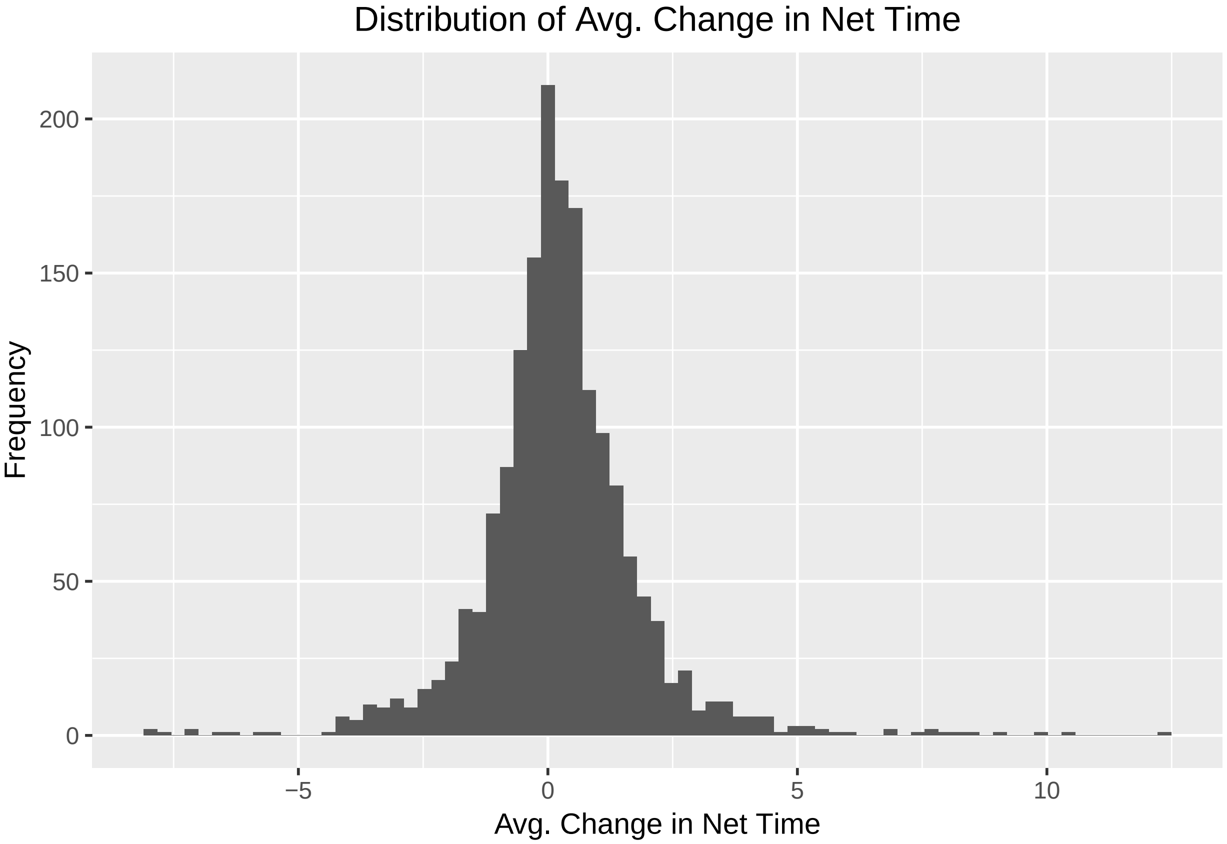
III. Results

The slopes from the linear models created for each individual model's net time over 7 years were effectively the average change in net time for that individual year to year. When we did this for all the runners, was initially surprising about these results was that the mean (0.2609) and median (0.1650) of the average changes in net time were both positive, meaning instead of improving over time, the average runner became slower. However, a counterargument that could possibly be used to explain this phenomenon is the fact that runners are obviously aging throughout this time period as well. 7 years is a long time and unless one is a professional runner or 18 years old when running the New Haven Road Race for the first time, it is unlikely that by the end of 7 years, one's body is as fit (let alone more fit) as it was at the start.

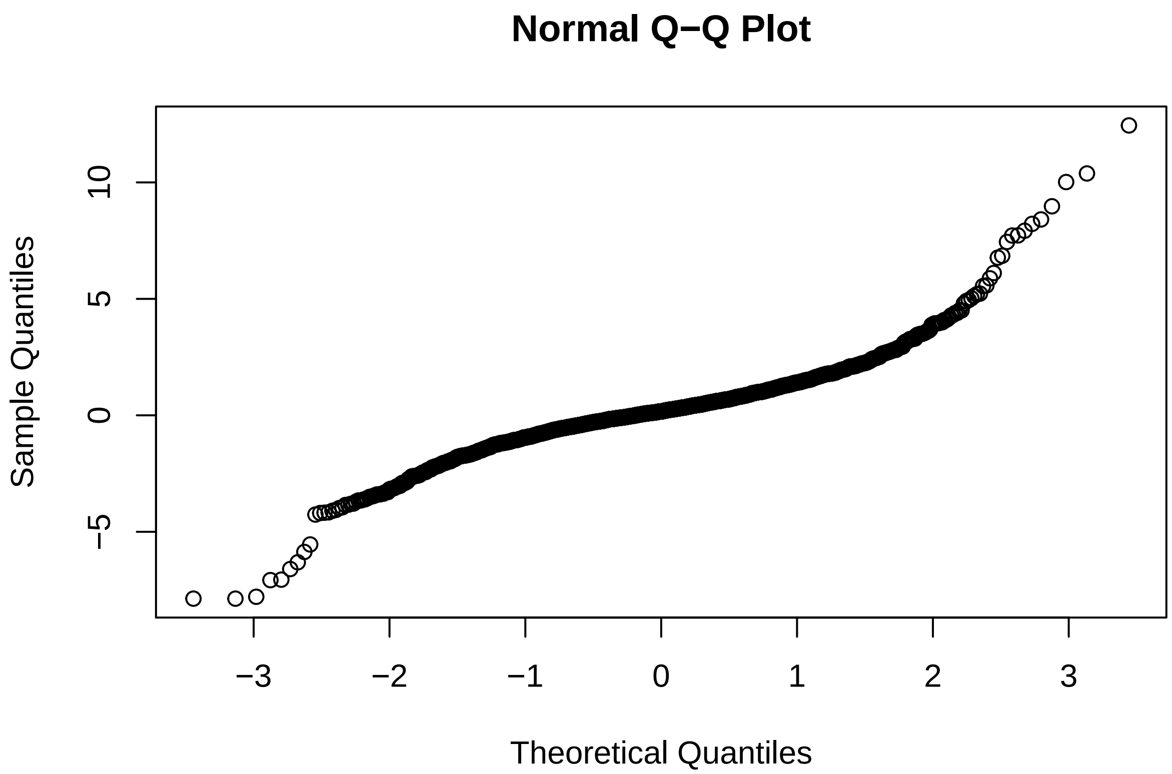
To model the situation, we proposed a normal distribution with unknown mean and variance that average yearly changes per runner come from. That is,

$$x_i \overset{iid}{\sim} N(\mu, \sigma)$$

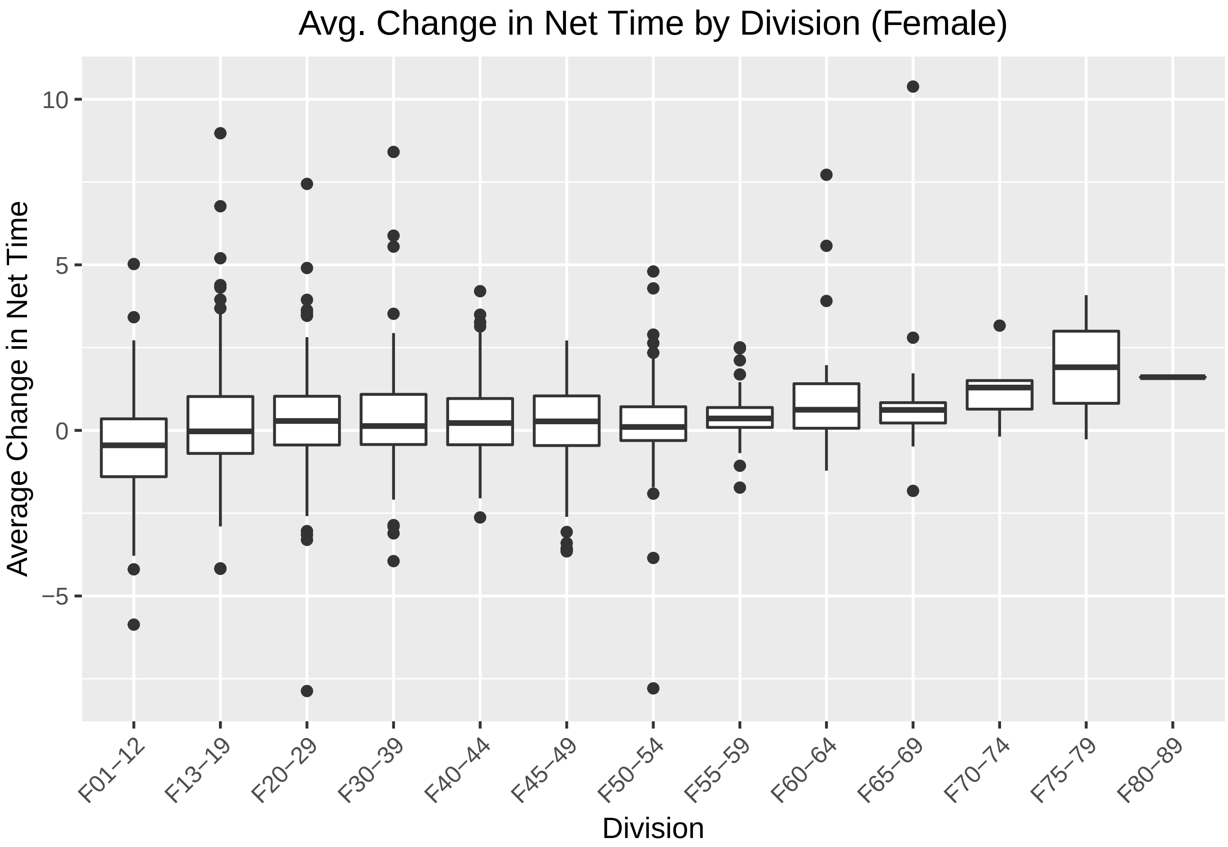
Thus, a two sided t-test with the null hypothesis being that the average change in net time year to year is zero, ($\mu = 0$), was used to see whether our results were significant or not. Despite a wide distribution, the average was conclusively greater than zero ($p < 10^{-8}$).

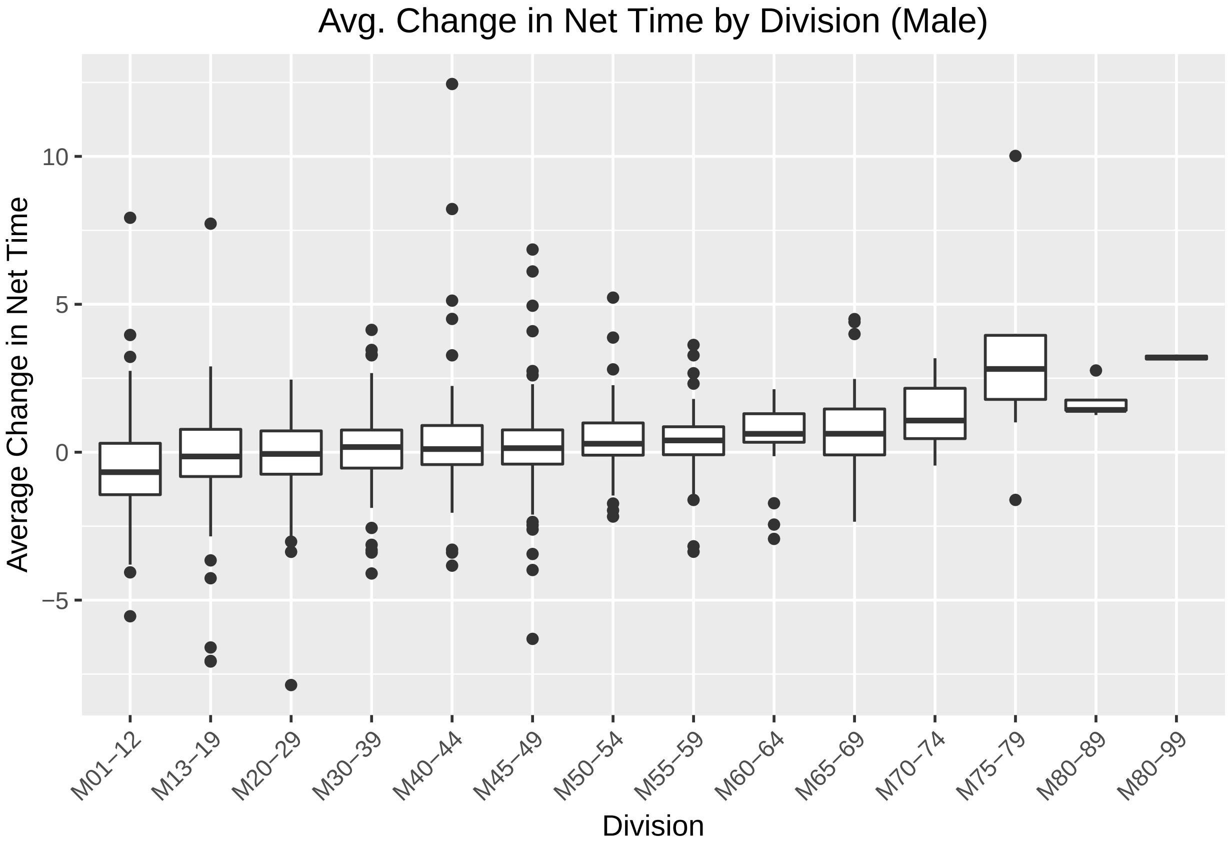


So according to our t-test, we would conclusively reject the null. However, the results of this t-test must be taken with a grain of salt. Although the distribution of the average changes in net times looks roughly symmetrical and bell-shaped at a glance, a qq-plot reveals that it is not quite normal, thus weakening the optimality of the t-test.



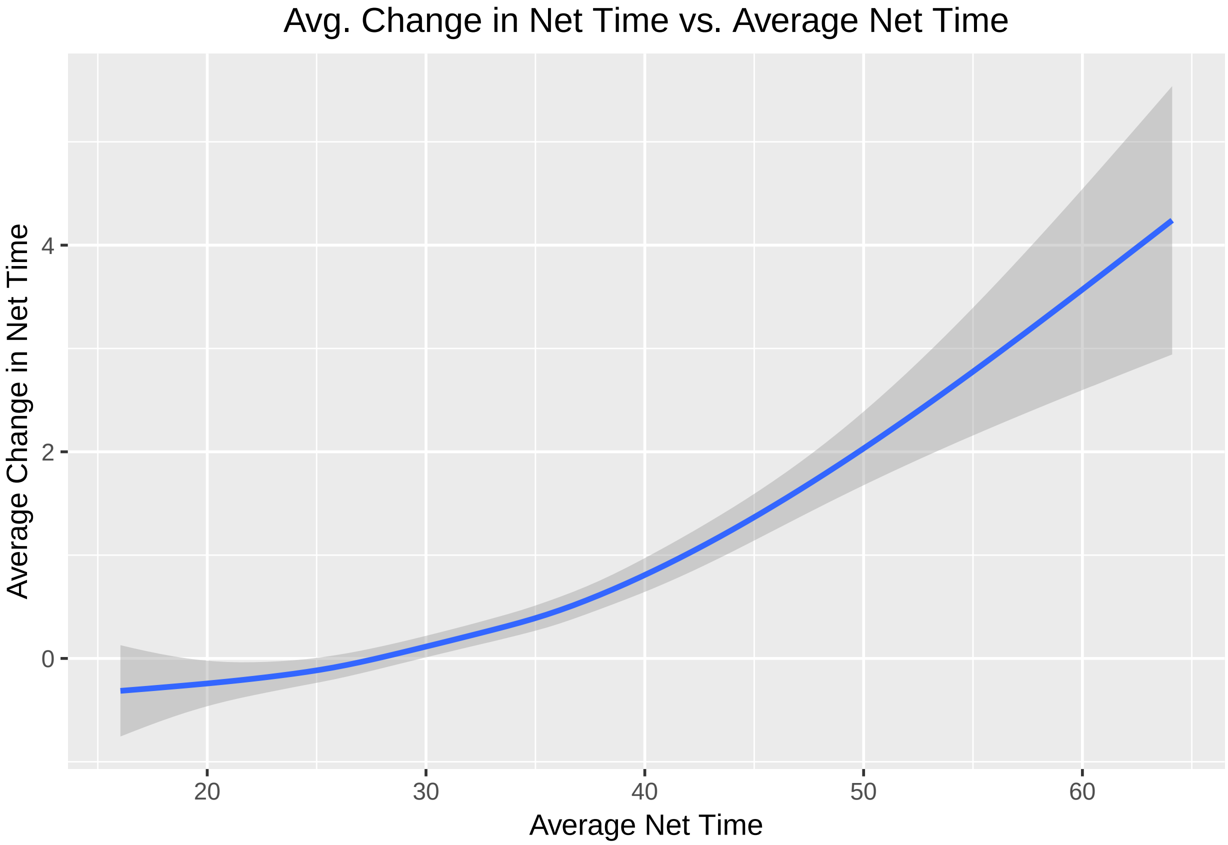
Nonetheless, the fact that the data revealed such a strong increase in net times over the years is quite curious. Following our previous line of reasoning that this could potentially be explained by the fact that runners are getting older with time and thus naturally becoming worse runners, we looked at the results when the data was split by division (as well as gender).


Avg. Change in Net Time by Division (Female)
Average Change in Net Time
10
5
0
−5
F01−12
F13−19
F20−29
F30−39
F40−44
F45−49
F50−54
F55−59
F60−64
F65−69
F70−74
F75−79
F80−89
Division


Avg. Change in Net Time by Division (Male)
Average Change in Net Time
10
5
0
−5
M01−12
M13−19
M20−29
M30−39
M40−44
M45−49
M50−54
M55−59
M60−64
M65−69
M70−74
M75−79
M80−89
M80−99
Division

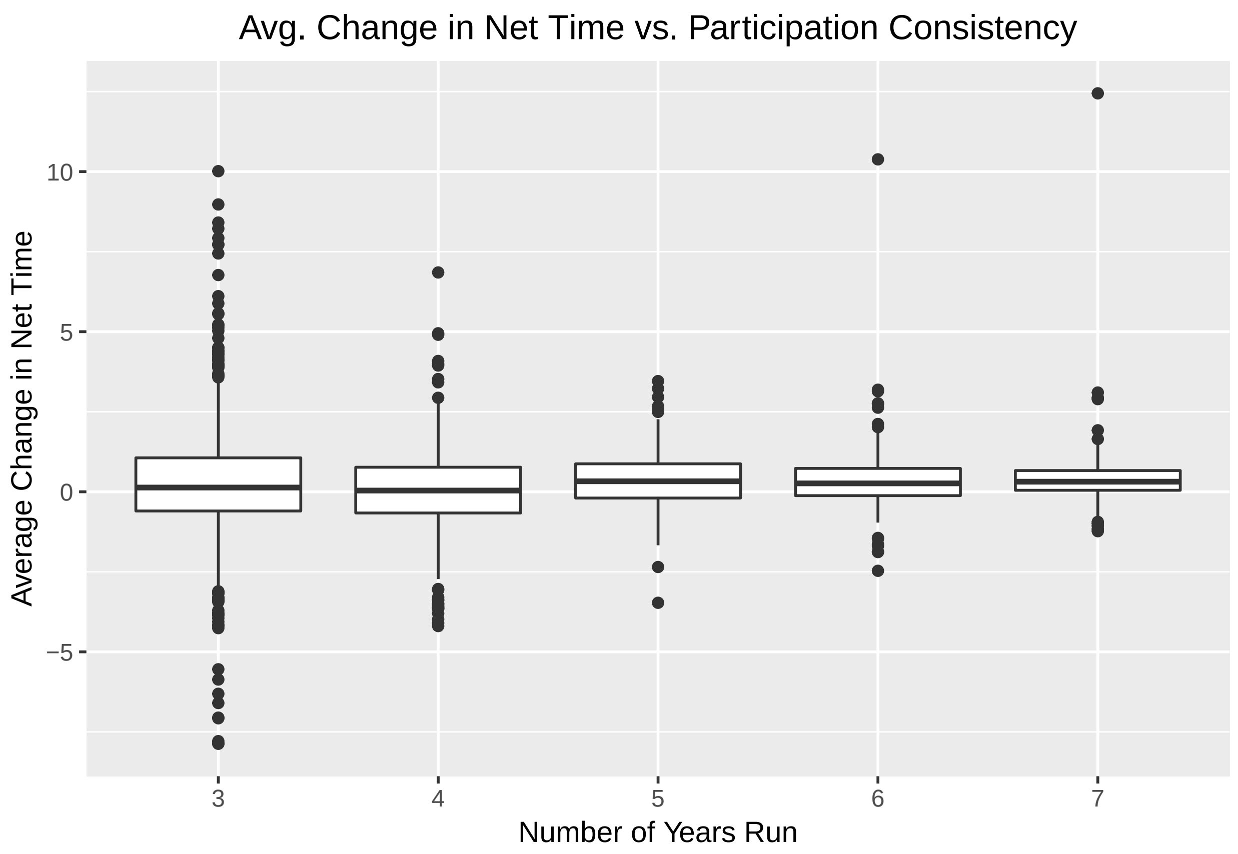
From both of these series of boxplots, we can see that the average change in net time year to year for a runner increases as we go up the divisions. In other words, on average it seems that older runners become slower more quickly. This seems to agree with our line of reasoning.

Furthermore, something that we initially expected to affect average change in net time was how fit the runner was on average. The best proxy available to express fitness was the average net time.



Fascinatingly enough, our results did indeed indicate that on average, slower runners faced greater average deterioration in net time over the years.

One final consideration was the effect of devotedness to this 5K race. We predicted the runners who ran more frequently in the race might be more avid runners, and might worsen less each year. This hypothesis was not supported in the plot, and statistical analysis found that no effect in this neighborhood was significant.



IV. Discussion

Even though our initial findings contradicted our preconceived beliefs that runners would improve in speed over the years (assuming they ran multiple times), there were some very interesting findings that could be used to explain why this was the case. Two trends that we managed to uncover were that age seemed to have a pretty strong negative impact on performance deterioration (runners in older divisions faced greater average increases in net time) and that average performance had a strong impact on performance deterioration (worse runners got even worse over the years on average).

The New Haven 5K Road Race includes many casual runners, who might not be as consistent year by year versus more dedicated ones. To attempt to validate this in our dataset, we took the number of years a runner had run in our NHRR 5K sample as a heuristic for being dedicated. Either the heuristic was faulty, or the effect of keeping up running is not large enough to overcome factors such as age.

Future work would include regressing on more variables, such as the weather at the time of the race. Stricter analysis would do more to analysis aggregate effect by year of the type “bad weather made runners in general slow during this year.”

Appendix

```
knitr::opts_chunk$set(echo = FALSE, results='hide', message=FALSE, tidy.opts=list(width.cutoff=80),tidy=
# Packages
library(ggplot2)
library(dplyr)
```

```r
x <- read.csv("NHRR_5K_Merge_ML.csv")
x$Name <- as.character(x$Name)
# Finding a Birth Year For Everyone
x$BirthYear <- x$Year - x$Age
# Finding intersection of rows with existing birth year and without existing birth year
names_by <- as.data.frame(intersect(unique(x$Name[!is.na(x$BirthYear)]), unique(x$Name[is.na(x$BirthYear
colnames(names_by) <- 'Name'
names_by$Name <- as.character(names_by$Name)
# Stripping x into only dataset with existing birth years
x_by <- x[!is.na(x$BirthYear),]


getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

for (i in 1:nrow(names_by)) {
  x[x$Name == names_by$Name[i],'BirthYear'] <- getmode(x_by$BirthYear[x_by$Name == names_by$Name[i]])
}

# Now that the ages that could be inferred have been inferred, this statement should be true - it will
stopifnot(length(intersect(unique(x$Name[!is.na(x$BirthYear)]), unique(x$Name[is.na(x$BirthYear)]))) ==

# Calculating Age
x$Age <- x$Year - x$BirthYear
x <- x[colnames(x) != 'BirthYear']


# Inferring Gender From Division
x$Sex[is.na(x$Sex)] <- substr(x$Div[is.na(x$Sex)],0,1)

# Eliminating the rows that have M01-19 or F01-19 as a division
x <- x[!((x$Div == 'M01-19' ) | (x$Div == 'F01-19')),]

# Keeping only people who run more than once
reprun <- x[x$Name %in% unique(x$Name[duplicated(x$Name)]),]
reprun <- reprun[!is.na(reprun$Div),]

# Eliminating the People with Same names that ran the same year
library(tidyr)
library(dplyr)

reprun$name_year <- paste(reprun$Name, reprun$Year)
reprun <- reprun[!(duplicated(reprun$name_year) | duplicated(reprun$name_year, fromLast = T)), ]
reprun <- reprun[colnames(reprun) != 'name_year']
write.csv(reprun, file = "NHRR_5K_Merge_Clean.csv", row.names = FALSE)
df <- read.csv("NHRR_5K_Merge_Clean.csv", as.is=T)
divs <- df %>% distinct(Div)
divs[order(divs),]

# Arrange by year and nettime
df %>% arrange(Year) %>% select(Year, Nettime)
```

```
# only runners who've run every year
superRunners <- df %>% group_by(Name) %>% filter(n() == 7) %>% select(Name, Age, Div, Year, Nettime, Se
# Looking at Age vs. Nettime for Everybody
df %>% ggplot(aes(x=Age, y=Nettime, color=Sex)) + geom_smooth(na.rm = T) + theme(plot.title = element_t
# superrunner individual performances over time
set.seed(123)
names <- sample(superRunners$Name, 10)
superRunners[superRunners$Name %in% names,]
superRunners[superRunners$Name %in% names,] %>% ggplot(aes(x=Year, y=Nettime, col=Name)) + geom_line()

# lm on yearly difference for runner
andrea <- df[df$Name == "Andrea Benjamin",]
# andrea$Year <- andrea$Year - 2012 # not needed; affects the p-val of intercept but not slope of bestf
m <- lm(Nettime ~ Year, data=andrea)
m$coefficients[["Year"]]
summary(m)
lm(Nettime ~ Year, data=andrea)$coefficients[["Year"]]
lm(andrea[["Nettime"]] ~ andrea[["Year"]])
# All together
#superRunners$Year <- superRunners$Year - min(superRunners$Year)
supRunChanges <- superRunners %>% group_by(Name) %>% summarise(effectDueToYear = lm(Nettime ~ Year)$coef
# df %>% group_by(Name) %>% filter(n() == 2)
runChanges <- df %>% group_by(Name) %>% filter(n() > 2) %>% summarise(effectDueToYear = lm(Nettime ~ Yea
summary(runChanges$effectDueToYear)
runChanges[runChanges$effectDueToYear<20,] %>% ggplot(aes(x=effectDueToYear)) + geom_histogram(bins=75)
t.test(runChanges$effectDueToYear)
qqnorm(runChanges[runChanges$effectDueToYear<20,]$effectDueToYear)
women <- runChanges[runChanges$Sex=="F",]
ggplot(women, aes(x=Div, y=effectDueToYear)) + geom_boxplot() + theme(plot.title = element_text(hjust =
men <- runChanges[runChanges$Sex=="M",]
men <- men[men$effectDueToYear < 20,]
ggplot(men, aes(x=Div, y=effectDueToYear)) + geom_boxplot() + theme(plot.title = element_text(hjust = 0
runChanges_2 <- runChanges[runChanges$effectDueToYear <20,]
ggplot(runChanges_2, aes(x=meanNettime, y=effectDueToYear)) + geom_smooth() + theme(plot.title = elemen
ggplot(runChanges[runChanges$n<=7 & runChanges$effectDueToYear<20,], aes(x=as.factor(n), y=effectDueToYe
```