Cherry Blossom 10 Mile Run

Michael Wells, Justin Valentine, Mina Mehdinia

December 07, 2022

Abstract:

In this project, we analyze the data from the Cherry Blossom 10-mile running race from 1973-2022, which took place in Washington DC, in order to elucidate the connection between age and physical fitness with respect to how fast people run. By using R, we processed the data and we found some interesting results about participants.

Executive Summary:

We found a negative relationship between age and health. We also discovered that running time deteriorates faster as age increases. Higher temperatures resulted in higher running times. There was a negative relationship between precipitation and running time; that is, higher precipitation coincided with lower running times. Most runners came from the eastern United States.

Introduction:

The goal of this project is to use the Cherry Blossom data to draw conclusions about health. Specifically, we want to analyze how age affects health. We will do this by looking at how age affects a runner's time in the Cherry Blossom race. A slower time indicates poorer fitness. We hypothesize that age has a negative effect on health. We also wish to explore whether the negative effect is magnified in the higher range of ages. Another thing that we will look at is which states in the USA had the greatest share of runners in the Cherry Blossom race. This will let us restrict our conclusions to the states with the greatest number of runners. In addition, we will look at how temperature and precipitation affect running times. Our hypothesis is that higher temperatures and greater amounts of precipitation will result in higher running times.

We find that indeed age does have a negative effect on health and that the effect is greater for higher ages, especially in men. We find that higher temperatures result in higher running times; however, we were surprised to find that higher rates of precipitation resulted in lower running times. Our hypothesis for this paradoxical result is that precipitation has a cooling effect, which enables runners to run faster.

Background:

The Cherry Blossom 10 mile run is a popular race that people use to train for the Boston Marathon, and a tourist friendly run for visitors during the Cherry Blossom season. The data is available from a publicly available website here: https://www.cballtimeresults.org/performances. Participants were required to finish in under two hours and twenty minutes. The range of ages attending this race is 9-84. We needed to scrape this data from the website, edit (clean) data values which were recorded incorrectly, remove missing values, add additional information (re: weather), and analyze it for trends using regression models and plots.

Data:

Every data point represents a runner's participation in the 10 mile run, and includes some information as described below.

- Race: The year of the event
- Name: Runners first and last name
- Age: The age of the runner at the time of the race

- Time: The completion time of the runner's race
- Pace: Defined as time per mile
- Pis.Tis: Position in sex out of total in sex
- Division: runners sex and age group
- PiD.TiD: Position in division out of total in division
- Hometown: Runners hometown

The data is only available right now on that website, therefore we needed to scrape it to be usable in R. The first step we did was using the "rvest" and "tidyverse" libraries, and function read_html() to scrape the data from the website. There were 554 pages for each sex category(women and men) so we wrote a for loop that goes through each page and also each year(1973-2022). After we collected that, we converted the data to a data frame using as.data.frame(). The dimension of the dataset after scraping it from the website was 181614 rows and 9 columns for the women's data set and 188438 rows and 9 columns for the men's dataset.

Data Cleaning and Reformatting Variable:

- Convert Time

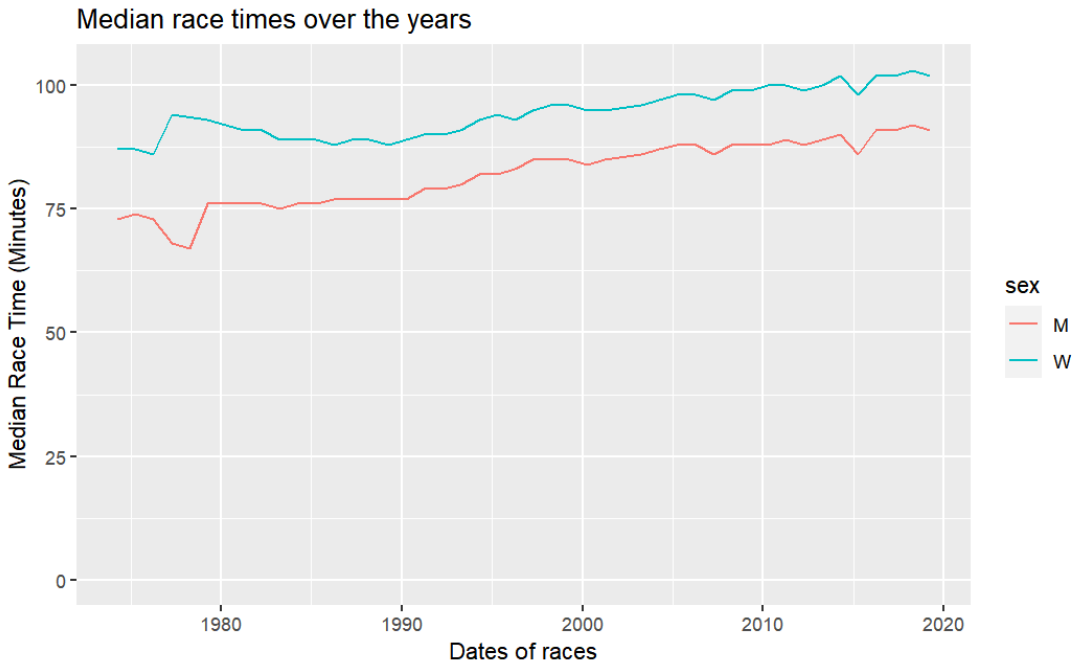
We want time in a numeric format so it can be more easily summarized and modeled. The first issue we encountered was differing formats for data entry. For example, a time of 56 minutes and 40 seconds could be recorded as 0:56:40 or 56:40:00. We filtered the data and, using str_detect, formatted all of the times the same way, and then converted to minutes. Times greater than 140 minutes were removed. Three values which were recorded as decimals were removed.

- Removing Missing data

The dataset had many missing values. Missing data is indicated by "NR", "NA", or just "". We first started with removing rows with NA in Age and Time observations for both the women's and men's dataset, since these two are the most important variables we are looking at. Then we removed the rows in which time was not recorded. Age was restricted to greater than or equal to 9.

- Cleaning Year

We noticed a significant dip in times for the year 2015.



With many of the top times in this year being supposed world-record times we thought there had to be serious issues with mis-recording times. After consulting the Rite-of-Spring document outlining events from each race it was revealed that this year was only a race of 9.39 miles.

2015 (April 12)

For the first time in its history, the Credit Union Cherry Blossom was not a 10 mile. Due to an accident on the course just 90 minutes before race time, organizers had to scramble to come up with an alternate route because of the accident investigation. The results showed some eye-popping times for anyone who did not read the explanation that the course was 9.39 miles long after the re-routing. Any disappointment over the shortened course was more than counterbalanced by the most

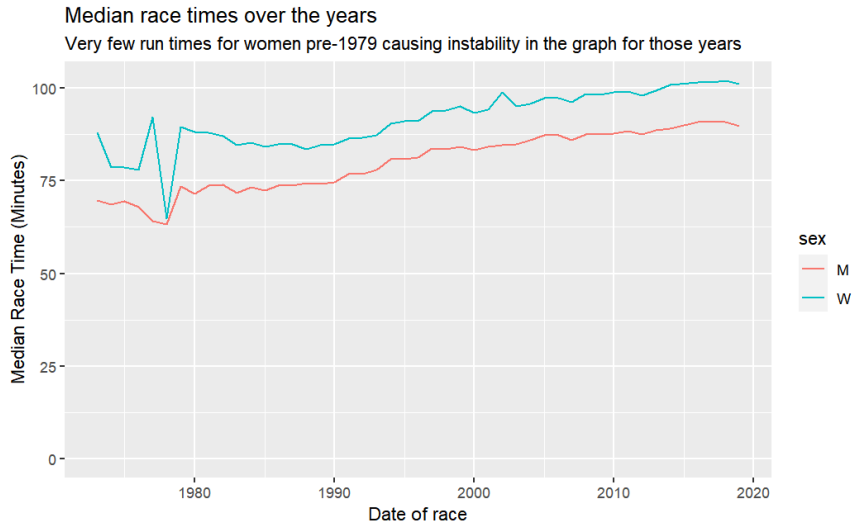
For this reason we chose to remove 2015 from the data set.
Additionally, after removing NA's from Age, we only had 2 data values for the year 1973, so all analysis that included Age and Year had this year removed as well.

- Cleaning Hometown

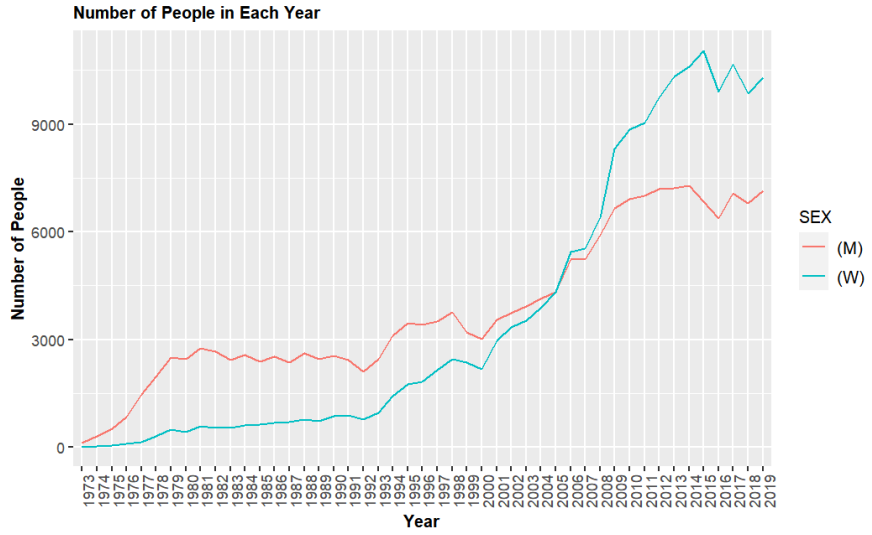
We wanted to see the number of runners from each state in the US. For that, we had to do some cleaning in "Hometown" in our dataset. First we remove the rows that didn't specify the runner's hometown. The format of the hometown is: city, State. Using the strsplit() function, we split the city and state and then keep only the state. When we dive into the dataset, we found out that some states have multiple different names. Therefore, we had to change them. For example for DC, we found: D.C, D.C., dc, d.c, d.c., District of Columbia, and the same for New York and Virginia.

Exploratory data analysis:

We wanted to see how the median race times changed over the years between men and women.

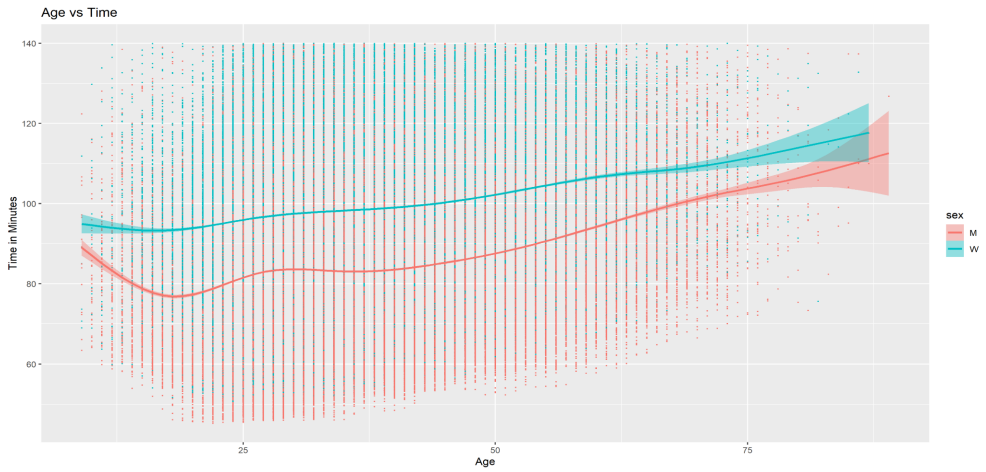


Prior to about 1980, the number of participants was small but grew rapidly (shown in next plot) which accounts for the instability in the median times there. As the years go on, the median times gradually trend upward, which we conclude are most likely the result of the increasing popularity of the race attracting more of the general public rather than mostly particularly skilled runners.

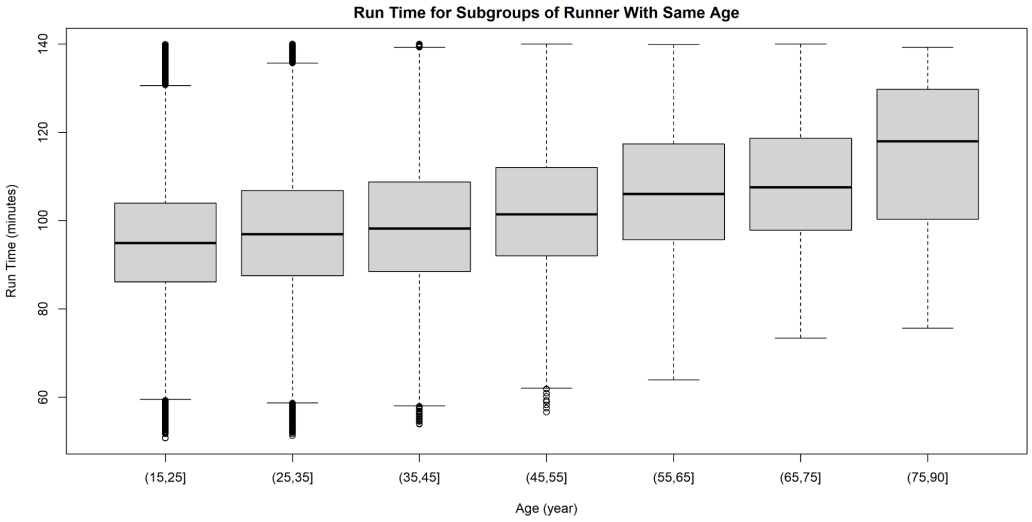


As shown, the number of male participants quickly grew until about 1978 then stayed pretty even until about 2000 when it quickly grew again. The number of women stayed fairly low until about 1993. 2005 was the first year that the women outnumbered the men and that has stayed consistent ever since.

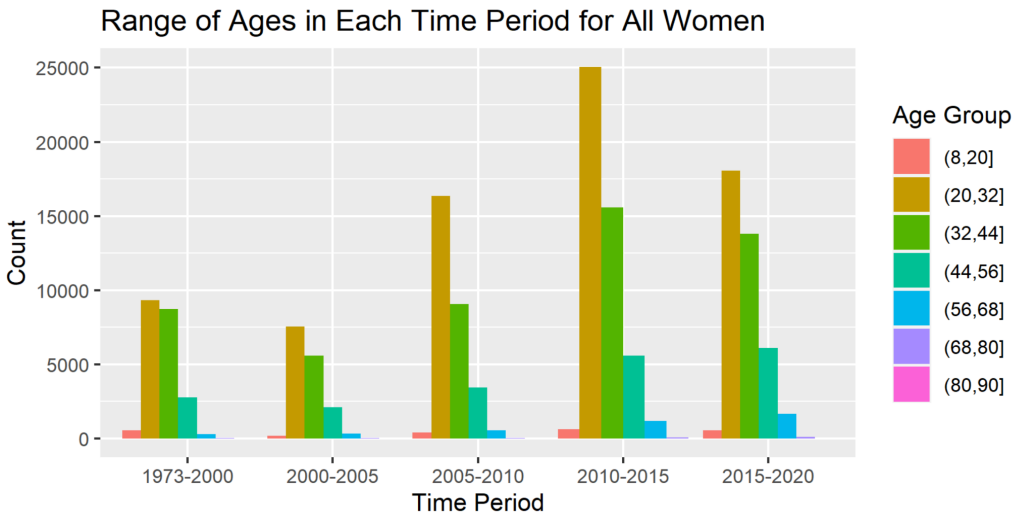
Another thing we wanted to see was the runner's age and the time they finish the race between men and women.



We can see that as a runner's age increases, time to finish the race increases too. Followingplot is also the plot of age v.s. Time but only for women.

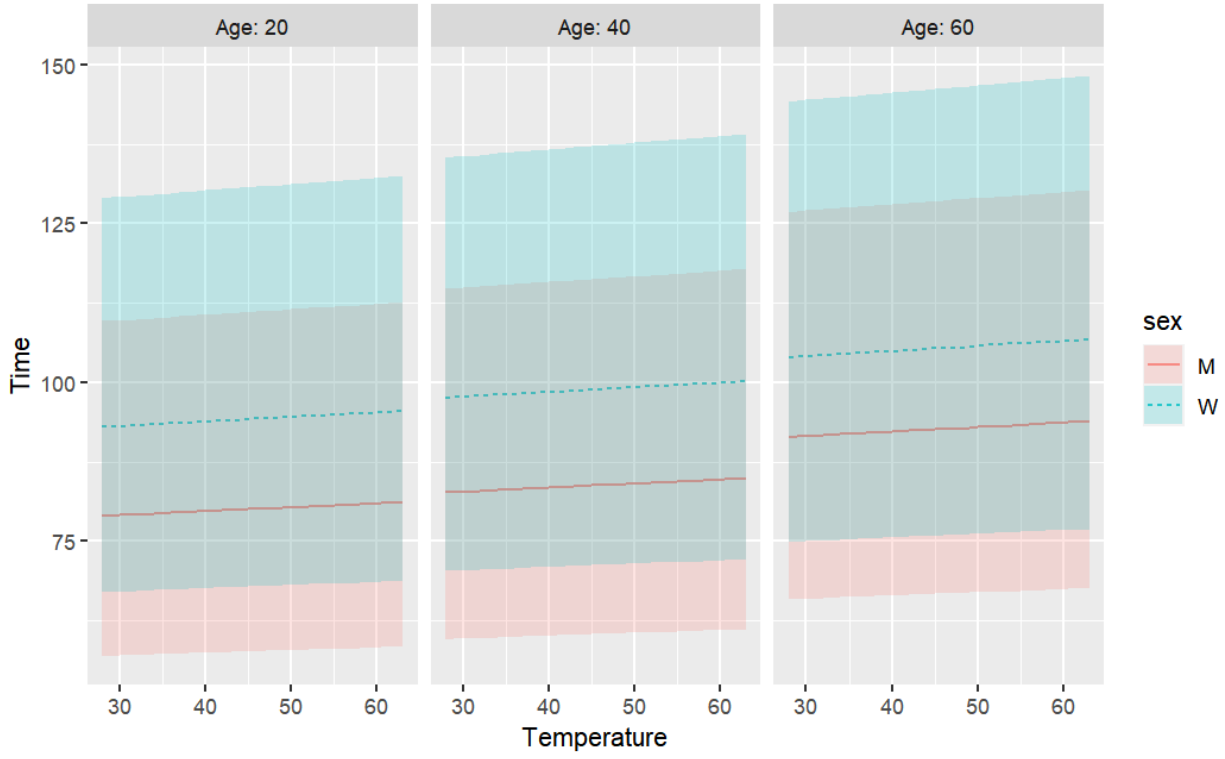


Here, we group the runners into a 10-years age interval and plot the summaries for each subgroup in the form of a boxplot. Similarly to the men's boxplot, we observe in this plot that the upper quartile increases faster with age than the median and lower quartile.
We were interested to see what the range of runners is in each time period. The age group we are interested in are defined by 12 range age.

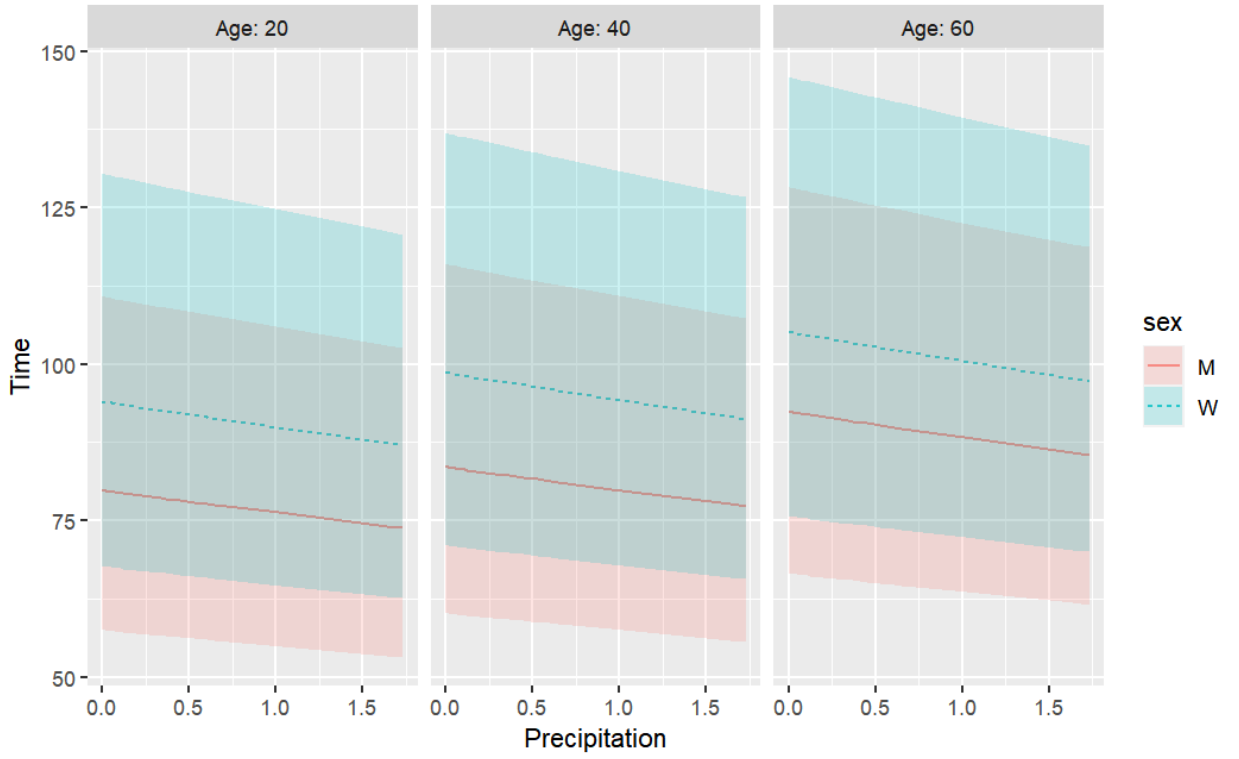


From the above plot, we can see that the age between 20 and 32 is the most in all year for Women. We have the same plot for men which was a bit different. It was indicated that the most age was between 32-44 in all time periods except between 2010-2015 which was age 20-32(see Appendix 2).

Another trend that we wanted to investigate was the effect of rain and temperature. This information was collected from NOAA and recorded at the National Arboretum in Washington DC. This information was analyzed in conjunction with a regression model (see Appendix 1).

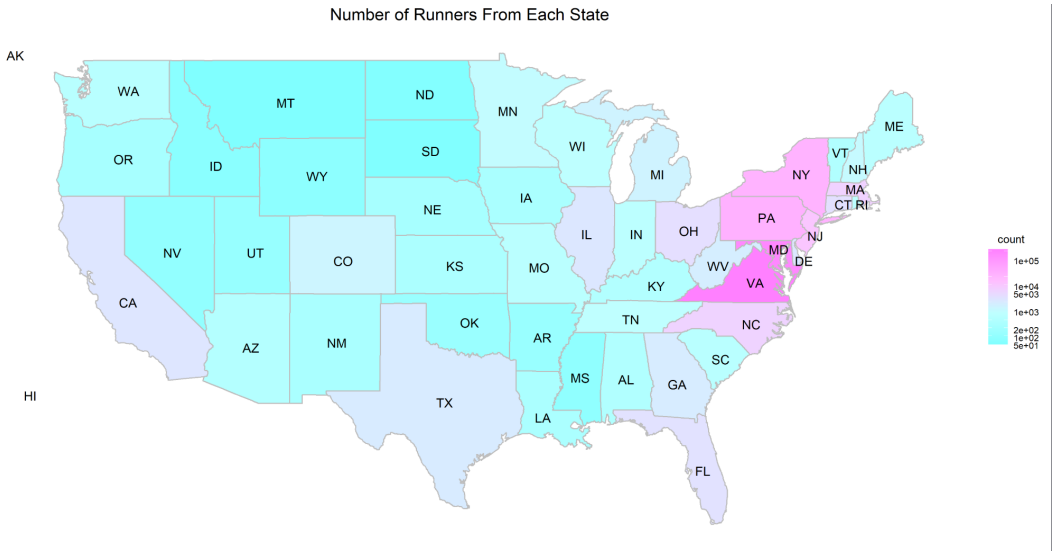


Here we can see the change in run times for men and women at 20, 40 and 60 years of age, for varying levels of the minimum recorded temperature that day (TMIN). There is a gradual increase in run time as the temperature increases from 30 to 60 degrees. The bands in pink and blue are the prediction intervals for the run times for each sex.



Similarly, we can see that as precipitation (PRCP) increases, run times across ages of 20, 40 and 60 years of age all decrease. This could be due to the rain having a cooling effect on runners.

Just for curiosity, we were wondering which states runners mostly come from in the US, and for that we did a map plot which indicates the number of runners in each state.



From the above plot we can see that runners mostly came from states: Virginia, Maryland, Washington DC, New York, New Jersey, Pennsylvania.

Methods:

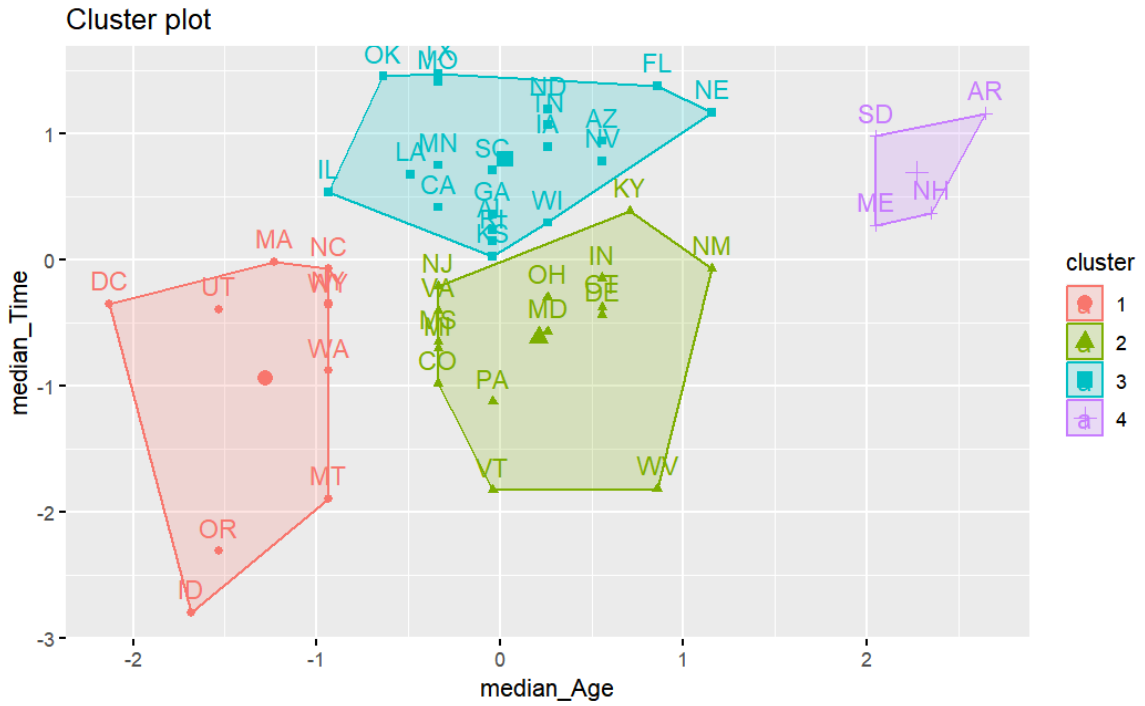
We modeled the data using a Linear Mixed Effects (LME) model on the Cherry Blossom data. An LME model consists of fixed effects and random effects. The fixed effects are like a general trend, while the random effects account for individual variation.
We ran this model on the subset consisting of runners who had run five or more races and had male gender (the exclusive focus on men was to simplify the data under consideration). This was a subset of the cleaned data used in our other methods. We assumed that individual runners would have times that were related to each other. We assumed that the age of the subject was related to their running time in a way that can be approximated by a line. Also, we assumed that the errors had constant variance and were normally distributed. This last assumption is technical and is required in order for the model to be reliable. See the Appendix for more technical details.
After fitting the model, we computed the average rate of change in running time for each runner. Our hypothesis was that these average rates of change will generally be positive. We assumed that this would indicate that age negatively affects running time, and hence health. We ran a Wilcoxon signed rank test on the slopes to test this hypothesis. This is a statistical test that can determine whether a set of numbers is “on average” positive or not.

Results:

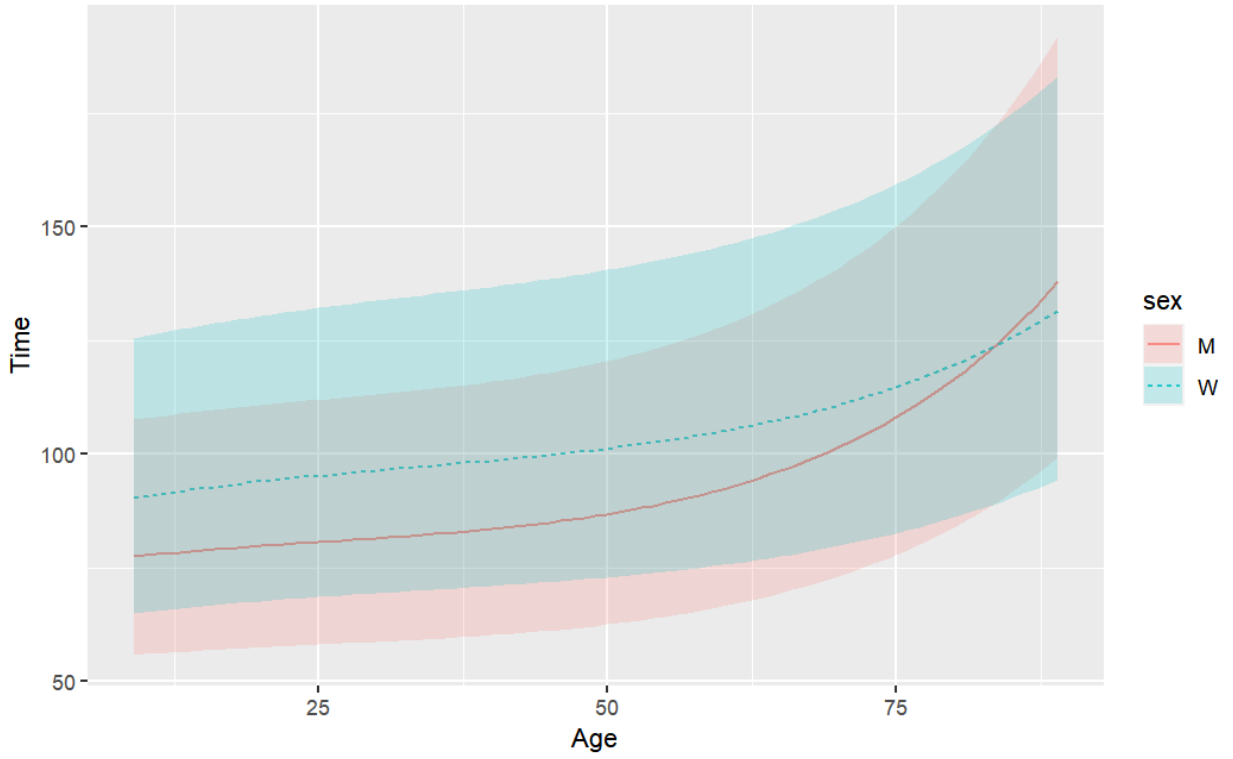
We got a marginal R^2 of 0.1475 and a conditional R^2 of 0.8828 for the LME model. These technical details show that the fixed effects do not explain the data very well, while the fixed effects combined with the random effects do a good job of explaining the data. We take this to mean that individual runners vary significantly in their performance. The Wilcoxon test showed that indeed the slopes were generally positive with high probability, confirming our hypothesis.

Discussion and conclusion:

We wanted to see if there were any similarities between runners from the various US states. We looked at the median age and median run time for everyone grouped by state and noticed they were clustered into distinct groups as shown below.



Using the regression model from Appendix 1, we were able to see how run time (Time.y) is predicted to change as Age increases.



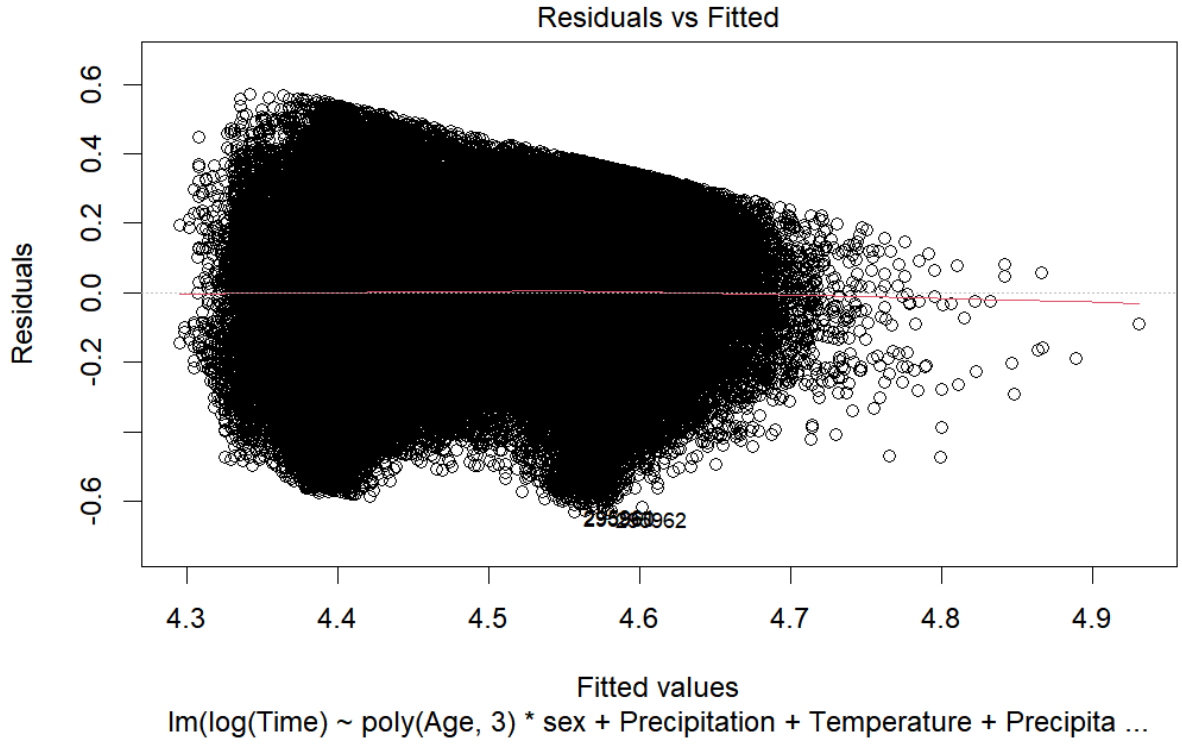
The increase is gradual for both men and women. Once men are around age 60 their times rapidly start to increase, while the increase for women is much more subtle.
The LME model with the Wilcoxon test showed a negative relationship between age and health. These results should be taken with caution, however, because several modeling assumptions did not hold: namely, that the errors should have constant variance and that the errors be normally distributed.
We recommend exploring further our finding that running times increase more rapidly after age sixty. The negative effects of age on health seem to get more pronounced as age increases.
A more complicated model may fix the issues with the LME, or perhaps transforming the variables in the model would be sufficient.

Appendix:

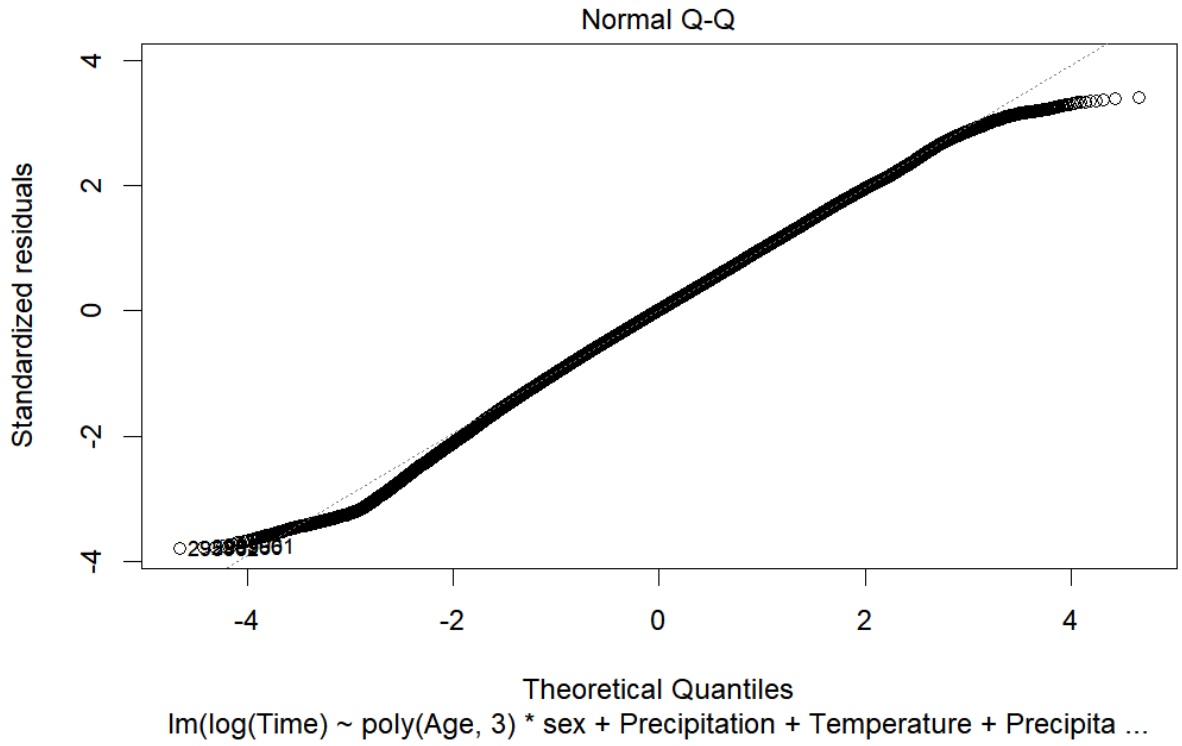
1.

The first regression model we used compared the run times as the response, against the interaction of sex and age, with age being raised to the 2nd and 3rd powers, and also the minimum temperature and amount of precipitation, temperature and precipitation interaction. Output is below.
Call:

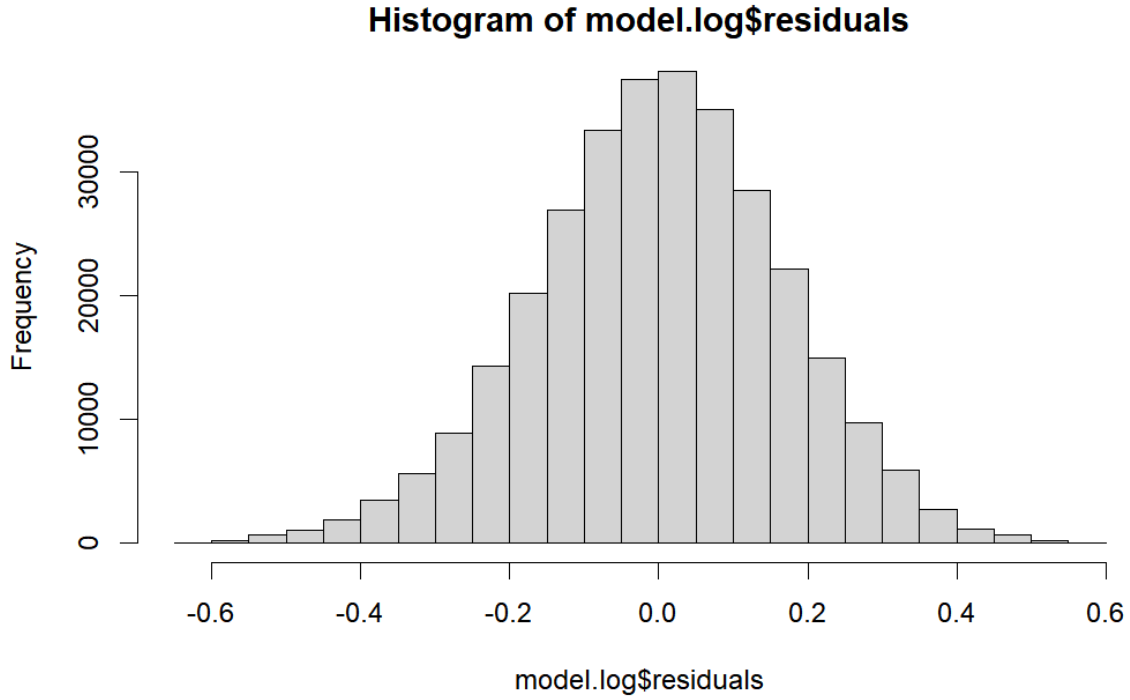
```
lm(formula = log(Time) ~ poly(Age, 3) * sex + Precipitation +
    Temperature + Precipitation * Temperature, data = data_w)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63430 -0.10820  0.00311  0.11223  0.56994

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.390e+00  2.312e-03 1899.21   <2e-16 ***
poly(Age, 3)1                2.020e+01  2.294e-01   88.08   <2e-16 ***
poly(Age, 3)2                6.017e+00  2.171e-01   27.71   <2e-16 ***
poly(Age, 3)3                2.107e+00  2.080e-01   10.13   <2e-16 ***
sexW                         1.622e-01  6.121e-04  265.04   <2e-16 ***
Precipitation                2.464e-01  9.777e-03   25.20   <2e-16 ***
Temperature                  7.751e-04  5.493e-05   14.11   <2e-16 ***
poly(Age, 3)1:sexW          -4.706e+00  3.525e-01  -13.35   <2e-16 ***
poly(Age, 3)2:sexW          -4.068e+00  3.565e-01  -11.41   <2e-16 ***
poly(Age, 3)3:sexW          -7.017e-01  3.581e-01   -1.96     0.05 .
Precipitation:Temperature -6.935e-03  2.466e-04  -28.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1673 on 313381 degrees of freedom
Multiple R-squared:  0.2035,  Adjusted R-squared:  0.2035
F-statistic:  8007 on 10 and 313381 DF,  p-value: < 2.2e-16
```



The residuals were impacted from above due to the times being restricted to below 140 minutes whereas there was no restriction on the lower end of the times.



The response was transformed by the logarithm of Time to try to improve the normality assumption.



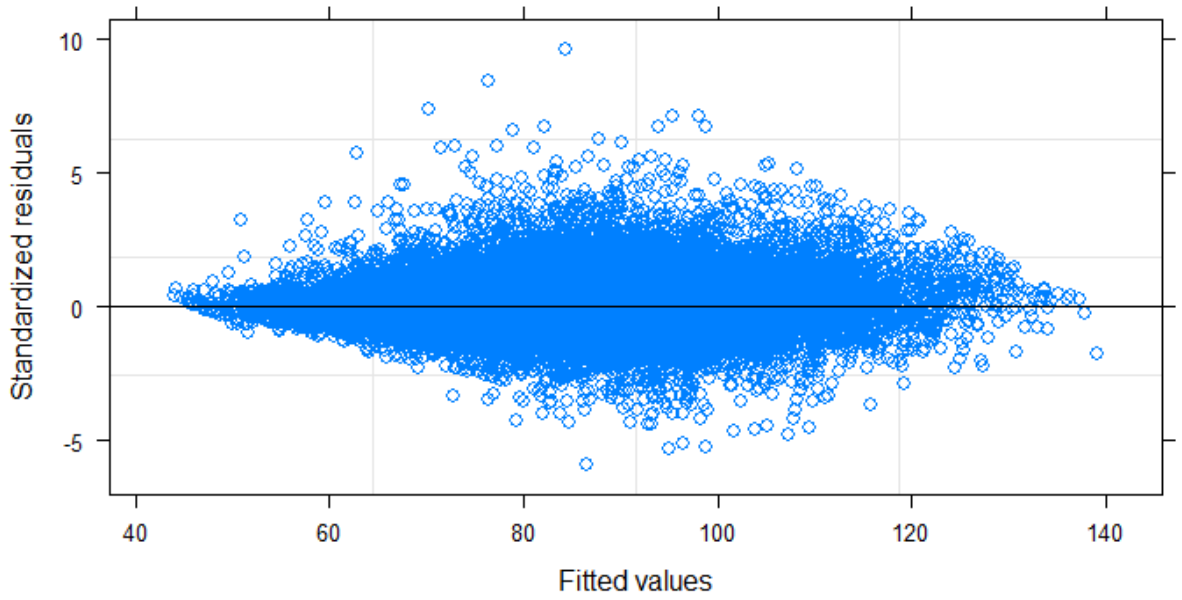
The next model we tried was a Linear Mixed Effects model with random slopes and random intercepts. We ran this on a subset of the data consisting only of runners who had run five or more races and also of male gender. The R package we used was nlme. The formula for the model was given by:

```
model = lme(Time.y ~ Age, data=dg,random=~1+Age | ID)
```

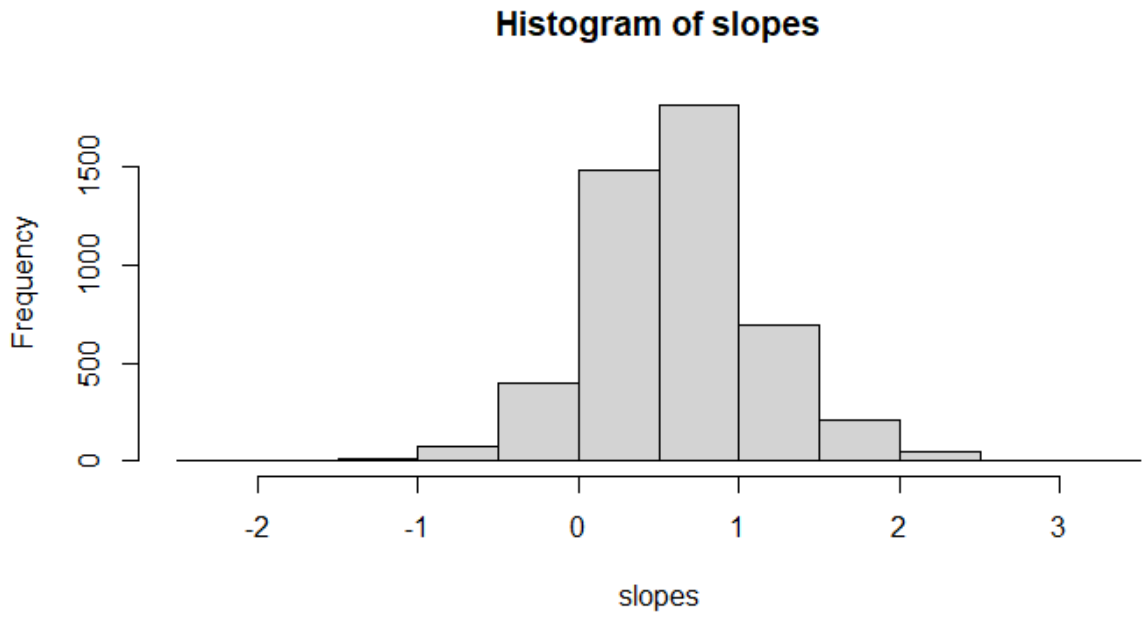
The output for the model was:

```
Linear mixed-effects model fit by REML
  Data: dg
        AIC      BIC      logLik
  241284.7 241335.4 -120636.3

Random effects:
 Formula: ~1 + Age | ID
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev           Corr
(Intercept) 32.7991103 (Intr)
Age         0.7153626 -0.93
Residual    5.7065495

Fixed effects:  Time.y ~ Age
            Value Std.Error          DF  t-value p-value
(Intercept) 54.14211 0.6201892 30068 87.29934     0
Age         0.61606 0.0138325 30068 44.53707            0
 Correlation:
            (Intr)
Age -0.953

Standardized Within-Group Residuals:
            Min               Q1       Med               Q3       Max
-5.93356428 -0.49101902 -0.08404679  0.39659805  9.63955192

Number of Observations: 34819
Number of Groups: 4750
```

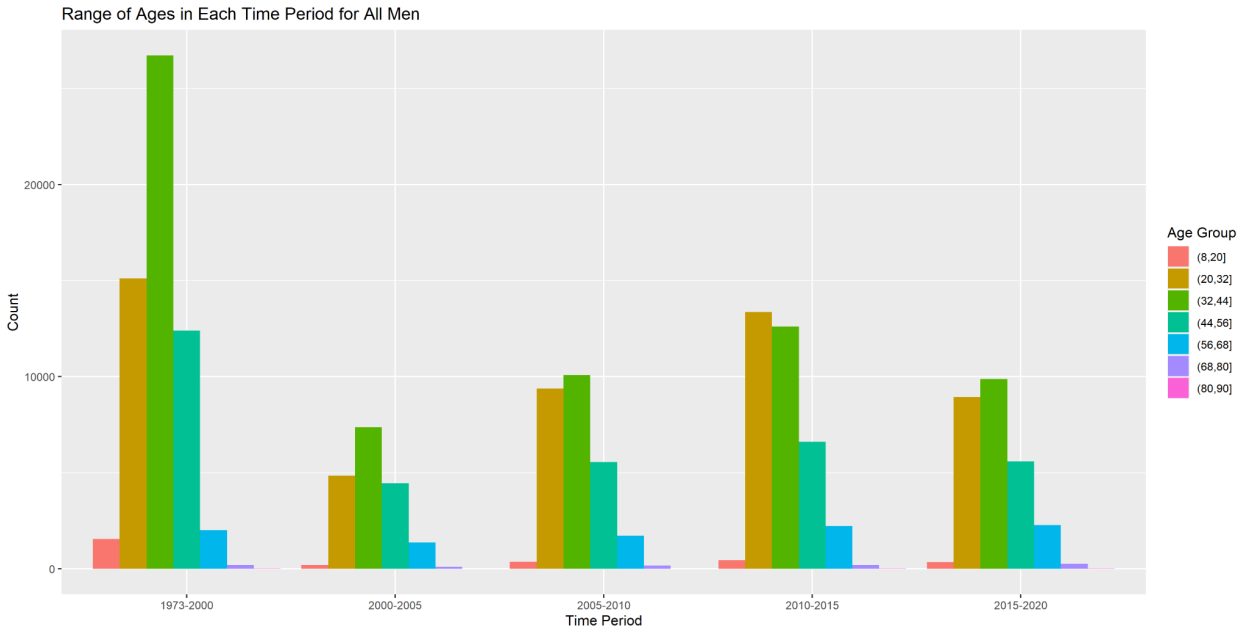
The residual plot is shown below:



As you can see, we don't have constant variance in the errors, as can be seen by the "football" shape in the plot. Constant variance would imply that the plot has no pattern. This is unfortunate, and the few transformations we tried on the variables didn't seem to help much. Also, the errors were not normally distributed. A normal distribution is the "bell-curve" distribution.
A histogram of the slopes computed for each individual runner is shown below:



2.



References:

-Various authors, 2022. “Mixed Model,” Wikipedia. url:

https://en.wikipedia.org/wiki/Mixed_model, Accessed: December 2022.