**Statistical Programming Final Paper**

Running has always been an interesting sport, mostly because the statistical revolution that has consumed baseball and other major sports has largely ignored it. In a way, running is in the same spot that baseball was some twenty years ago. Most coaches train from tradition or train based on body stats, like lactic threshold and VO2 max, instead of looking at the bigger picture and collecting larger amounts of data. Too much of the sport is based on a “feel”, getting the sweet spot between overwork and underwork. Even simple analysis for running is significantly lacking in size or scope, despite numerous research papers proving the merit of a data-based look at the sport.

Not only does the data-based approach interest us, the sport itself is interesting to us. Matt ran cross country and track all of his life, and still runs with the club team here at Virginia Tech. In addition, all of his family runs, so he’s not allowed to stop. Jack ran a 5k once in middle school. But other than teasing his roommate, who was a big runner in high school, about the complexity of the sport, he hasn’t had much recent experience with it. But when Matt explained how statistics could be applied to the sport, Jack became curious.

On top of the sentimental experiences, we already had an expansive dataset that covered 6 years of Virginia High School League (VHSL) cross country state meets, so we wouldn’t have to hunt down a cool or complete dataset from the dark corners of the internet, and since Matt used the data for a previous stat project, it would be easy to use and parse through because we knew the contents well.

So we basically wanted to ask; “How does the outside temperature affect running performances?” Everyone knows that running in the heat is not fun (some would say running in general isn’t fun), but running in extreme cold is just as bad, maybe worse, so how does the temperature affect this? Is there a “sweet spot”, a temperature range best for running?

To answer this question, we used the aforementioned state meet dataset. The dataset itself is populated with data from Athletic.net, in accordance to their terms and conditions (Definitely don’t want a cease and desist). In order to get all the data quickly and efficiently, a python web-scraper was created and customized for use on Athletic.net results pages. The scraper utilizes the BeautifulSoup4 and requests packages in order to get the website’s contents. For every site, both finisher and meet info is collected. For every meet the mean day temperature, date, and name is collected. A meet id is generated, consisting of the first letter of every word in the title and the last two digits of the year (“Clash with the Titans 2016” would become cwtt16). For every runner, their name, finish place, finish time, gender, and grade level is collected. Again, an id is generated, consisting of the first three letters of their first and last names and the first letters of their school (“Matthew Rogers” from “Hidden Valley” becomes matrogHV).

Once all of the data is scraped, it is printed to a csv using python’s CSV package. The meet data goes to a file called “meet\_info”, and the runner data goes to a file aptly called “runner\_info”. The scraper is general enough that any meet results page will work, but since it was created with the VHSL state meets in mind, there are some specific functions that get data for specific VHSL groups (1A - 6A), so more specific analysis can be done on this series of meets alone. For this project however, the only information we needed was the Daily temperature and meet ids from the meet\_info file, and all the finish times and meet ids from the runner\_info file.

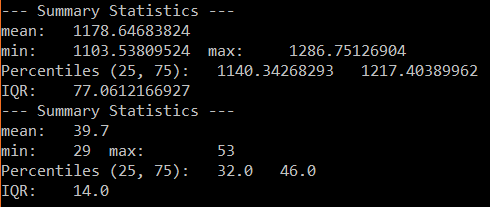
A few things to note about the data is that the meet takes place at the same time every year, so temperatures wont fluctuate a whole lot. I definitely thought about including meets earlier in the season (August/September), but a few problems would come with that. The first is that runners respond differently to course changes, so it would be difficult to conclusively attribute time disparities to temperature alone. The second is that early season races are slower by default because runners are not in top condition. To account for this, you’d have to make scales for season progressions, and my dataset is not large enough to do that.

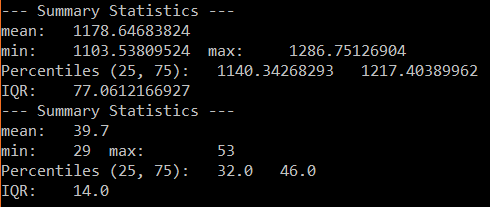
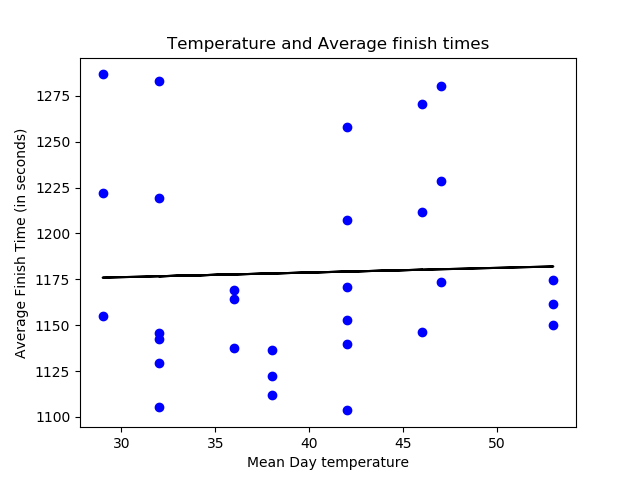
The methodology for answering our question was fairly straightforward. The first step was to read and merge the data, which we did with the help of our new favorite python package, pandas. Not only did this allow us to quickly and easily merge the data on a specific column, it was easier to visualize, as now we could run filtering commands similar to the ones you’d run in R, a language we were more familiar with.

The next step was to change the finish\_time column into a usable format. On import, the column is populated by strings in the format “M:S”. There are date-time functions built into pandas that would let us convert the column quickly, but we felt it would be easier to convert the column to seconds, so we built a function appropriately called “to\_secs”, that converts the column to total seconds.

Then, because we wanted an average of every race’s times, we made a for-loop to go through each of the group’s times on a per-year basis. The times and corresponding temperatures were stored in individual lists, and plotted using matplotlib. Because the lists came from a pandas dataframe, they had to be reshaped to a single list format. Neither of us are really sure why this had to be done, but it made the plot work. The line of best fit was made using the sklearn package. The r and r-squared values were procured the same way.

After making all that, we realized we didn’t have any summary stats, so we had to make a summary function that read a list and printed simple exploratory stats. It gives the min, max, the 25th and 75th percentiles and the IQR of the list of numbers given to it. Once we had all that run, and our graph made, it was time to look at our results.

Looking at our summary stats first, the variation in the temperature list is not as high as we might have wanted it, but seeing as this meet is at the same time every year, the variation is impressively high. The lowest temperature is a frigid 29 degrees fahrenheit, cold enough to make running fast very difficult. The highest temperature recorded is only 53 degrees fahrenheit, and while that is a little warm for genuinely fast racing, it isn’t anything sweltering. 

There is not a ton of variation in the finish times though. The IQR is only 77 seconds, so despite there being a relatively large variation in temperatures, there wasn’t that much of a variation in average race time. The scatter plot also shows essentially no correlation. The line of best fit is almost exactly flat, but tilts a little bit positive, which would mean that kids run slower as the temperature gets warmer, and in this case, warmer is high 40s-low 50s. Our r-value sits at 0.254, confirming the weak positive relationship seen in the graph. Therefore, this study showed that around these temperatures, race times and temperature have no correlation.

This study was good in the sense that there weren’t many confounding variables. The data we used was from the same race at the same course at the same time every year, and because it’s the state meet, the girls and guys field sizes are the same, meaning symmetrical data. But because this race is always at the same time of year, there isn’t enough variation in temperature to make a difference in race times. In the future, we might be able to take the same data from the state meet, but take data from different states. The Georgia state meet is a lot warmer than the New York state meet. This would give us a much larger variation in temperature. But as for this study, when the temperature is somewhere between freezing and 55 degrees fahrenheit, the race times aren’t affected by differences in the temperature in any meaningful way.