STATISTICAL ANALYSIS OF FRESH 15 RACE RESULTS

ALEX BEARDEN

1. Introduction

The FRESH 15 is a 15 kilometer road race held annually in Tyler, Texas. In this project, I scraped results from the nine-year history of the race and analyzed, both visually and via inferential statistical techniques, the relationships between gender, age, weather, overall time, and split differential (that is, how much faster or slower runners ran the second half of the race than the first). Most of the results were not surprising, but there were a few surprising and subtle relationships discovered in the analysis. The scripts and Jupyter notebook containing the analysis can be found at https://github.com/alex-bearden/race_results.

This type of race result analysis could be practically important in understanding the reasons that runners struggle in competition, especially if applied to longer races with more extreme conditions. (The main purpose of the present project, however, was simply to give me practice in obtaining, cleaning, and analyzing a real-world data set.)

2. Obtaining and Preparing the Data

The data was scraped from the web using a Python script employing the package Helium (which is an amazingly easy-to-use wrapper for Selenium—seriously, to get it to click the double arrow button to the next page, the line was literally: `click('>>')`; very little HTML/CSS parsing required). The scraped data was then checked for basic inconsistencies (such as missing rows or repeated rows) with another Python script.

The 9566 rows of data were combined with weather data retrieved from Weather Underground and loaded into a pandas DataFrame. The main relevant data preparation steps, which are described in more detail below, included adding the variables `Humidex`, `Half_split_differential`, and `Split_scaled`.

The humidex is a Canadian weather index that incorporates both air temperature and humidity, similar to the heat index used in American meteorology. The precise formula used in this project is the same as the one at this site:

$$\texttt{Humidex} = \texttt{Temperature} + 0.5555\left(6.11e^{5417.753\left(\frac{1}{273.15}-\frac{1}{273.15+\texttt{Dew\_point}}\right)}-10\right),$$

where `Temperature` and `Dew_point` are the air temperature and dew point temperature, respectively, in degrees Celsius. The reason I chose to work with the humidex rather than the heat index is that the usual heat index formula is not applicable for temperatures below 80°F, whereas the humidex formula seems to be applicable for lower temperatures. In our data, for example, the most recent two years had very similar temperatures, 61°F and 64°F, but a greater difference in dewpoint,

Date: August 16, 2022.

53°F vs. 60°F. As any runner who ran these two years of the race will attest, the difference in humidity was substantial and should certainly be accounted for. The humidex does a good job capturing this difference: it gives 65°F vs. 72°F.

One of the two dependent variables I wanted to analyze was the split differential, that is, how much faster or slower runners ran the first half of the race than the second. One issue with this in the data was that none of the years' results reported a halfway (7.5 km) split. (Most reported 5 km and 10 km splits, but I wanted a simple analysis of just one split variable.) The solution I used was to estimate half split differential with the formula

$$\texttt{Half\_split\_differential} = \texttt{Last\_5k\_split} - \texttt{First\_5k\_split},$$

which is equivalent to the difference between the natural estimates for first half split and last half split.

Finally, the `Split_scaled` variable was defined with essentially the following formula:

$$\texttt{Split\_scaled} = \texttt{Half\_split\_differential} \;/\; \texttt{Time},$$

where `Time` is the overall time for the race. The reasoning behind considering this variable is that it does a better job than raw split differential in capturing how well-paced a race is: a runner who runs the halves of the race in 60 minutes and 62 minutes has done a better job pacing, relative to his or her ability, than a runner who runs the halves of the race in 23 and 25 minutes. This was a good variable for visualization, but presented some problems in the regression analysis. Probably the most interesting part of this project to me was investigating the problems that arise when using a ratio built from two other variables in a regression analysis. This is discussed more in Section 4 below.

3. Visualizations

Here, we include some of the more interesting or illuminating visualizations of the data. See https://github.com/alex-bearden/race_results for many more.

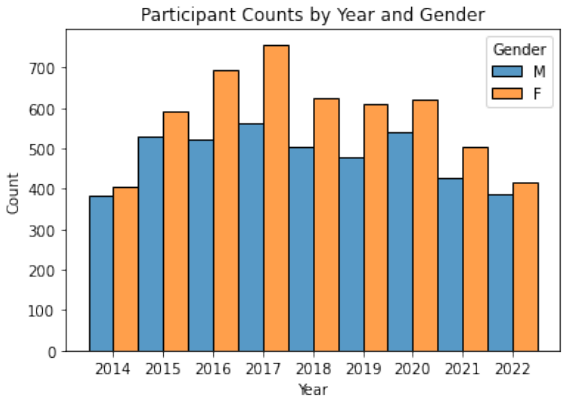


Figure 1. Note that the three years with the lowest counts were the unusual years: the first year and the two post-Covid years.

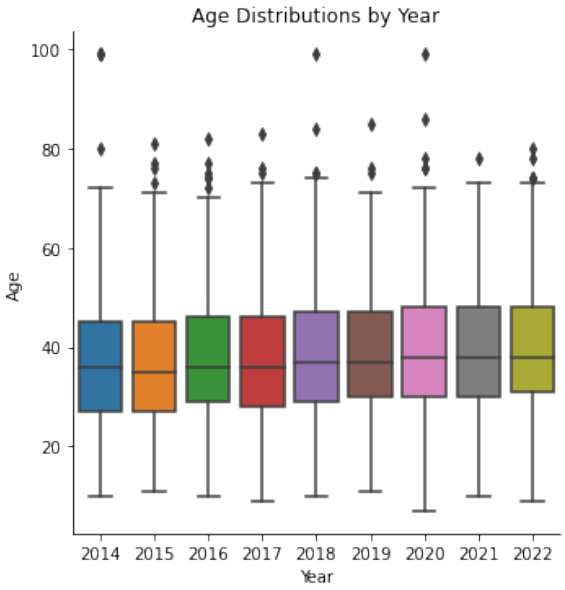


FIGURE 2. The age distributions of participants have steadily trended up throughout the years.

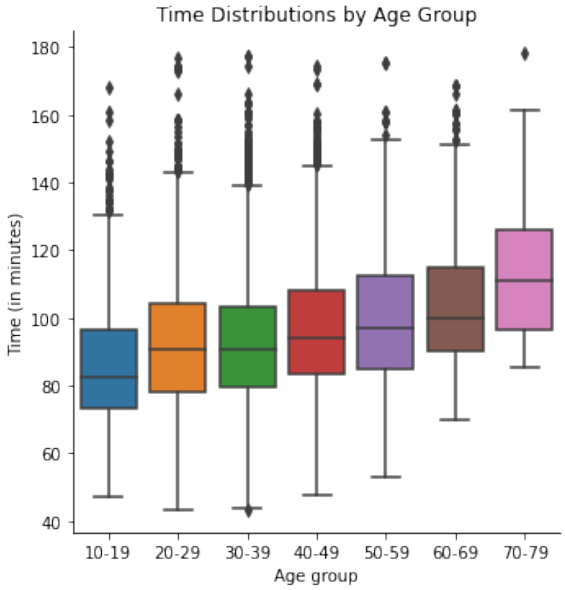


FIGURE 3. As a 34-year-old, I found the sameness of the orange and green box encouraging.

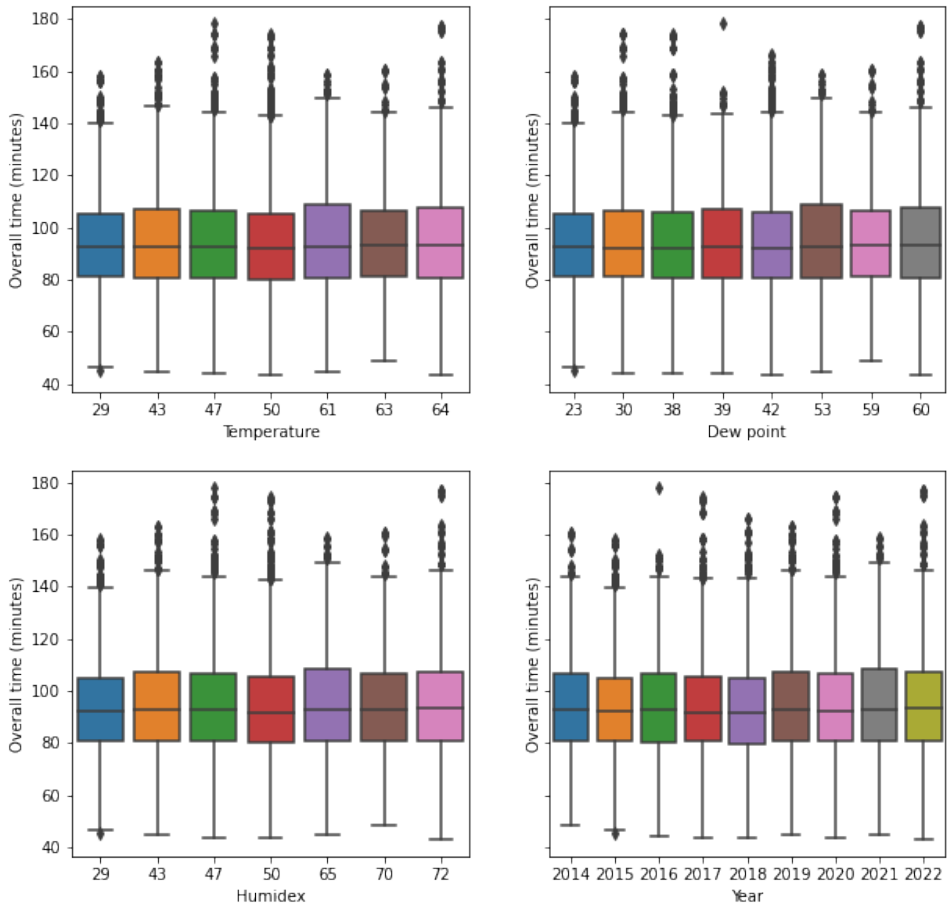


FIGURE 4. These were surprising to me: none of the weather data considered affected overall time much.

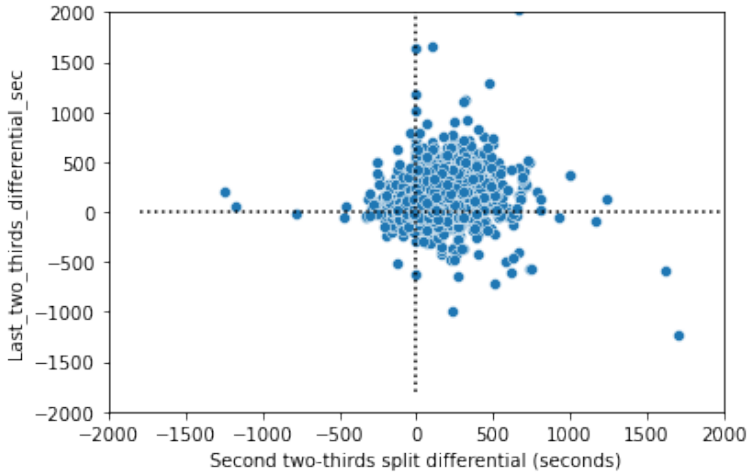


FIGURE 5. Each 5 km section of the race is more challenging than the previous, so unsurprisingly, most people (those in the upper right quadrant) ran each slower than the previous.

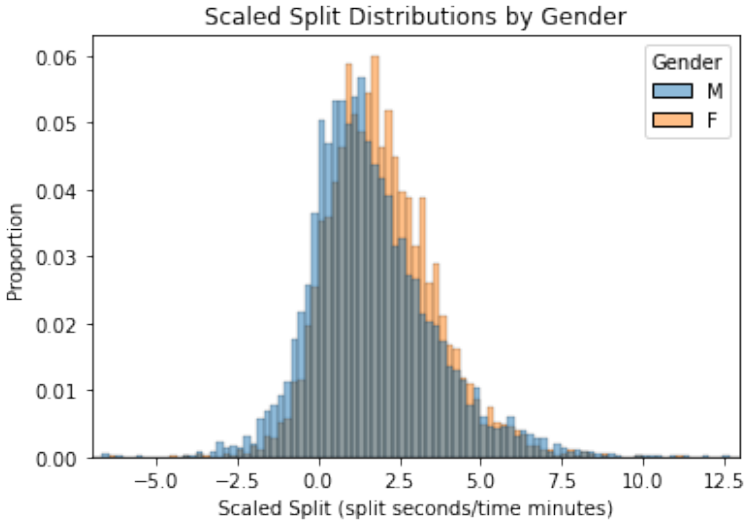


FIGURE 6. Even after scaling by time (which would artificially inflate females' split differentials), it appears that males run more evenly-paced races. As we discuss in Section 4 below, the truth is a bit more subtle.

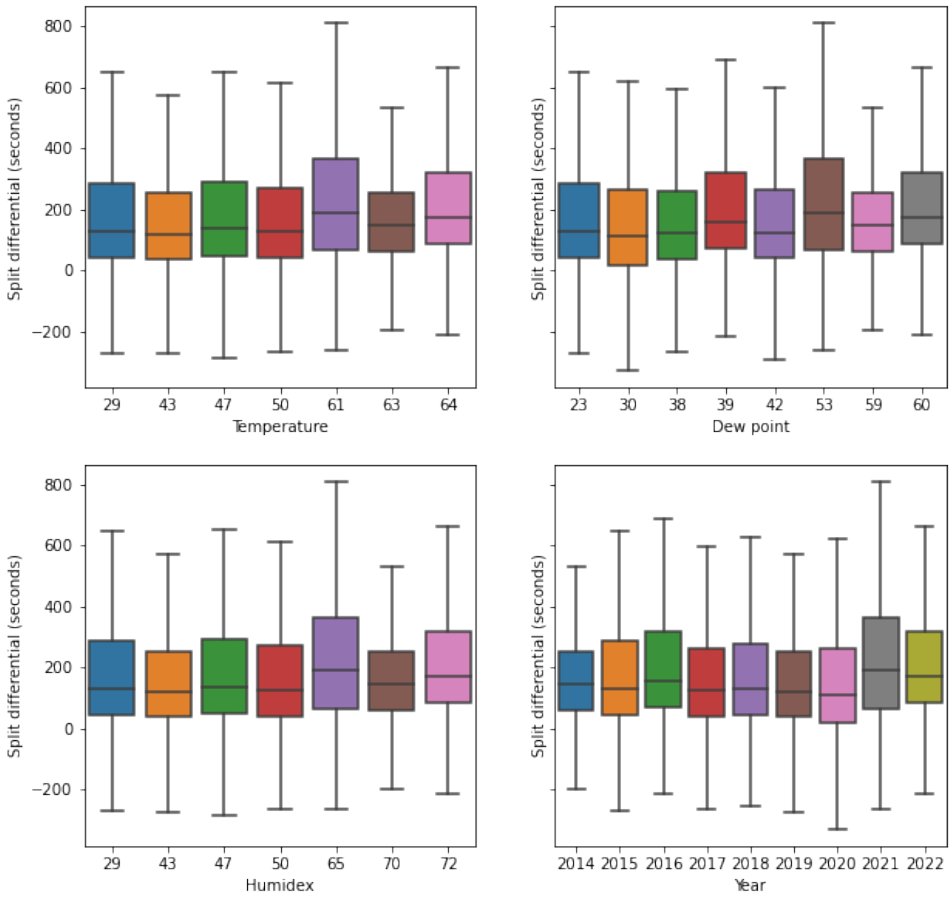


FIGURE 7. There's evidently a stronger positive relationship between weather and split differential than between weather and overall time.

4. Regression Analysis

The main goals with the regression analysis were to determine which variables out of `Gender`, `Age`, and `Humidex` were most predictive for `Time` and `Half_split_differential`, while controlling for `Time` with the latter. We used cross-validation to assess the predictive power of subsets of variables. The algorithm used to select the best subset was essentially the following:

(1) Split the data into ten folds.
(2) For each subset of predictors:
    (a) Train a linear regression model using these predictors on nine of the folds, and measure root mean squared error (RMSE) on the tenth fold.
    (b) Do this for each fold, then compute the average and standard deviation of the RMSEs over the ten folds as estimates of prediction error and its standard error (SE).
(3) List the subsets of predictors by estimated prediction error.

4.1. **Predicting Time.** As predictors for `Time`, the lowest error subset out of the choices `Gender`, `Age`, and `Humidex` was {`Gender`, `Age`}, and the smallest subset within one SE of the best was {`Gender`}. The main interesting conclusion from this, which also matches the visualizations above, was that weather, as represented by `Humidex` in this analysis, had no significant effect on overall time whatsoever. In fact, `Humidex` alone scored slightly worse than a constant predictor. Of course, this would certainly change if more extreme weather conditions were witnessed at this race: the warmest humidex in the dataset was 72°F (from a temperature of 64°F and a dew point of 60°F). Still, almost any runner would agree that a humidex of 72°F is warmer than ideal for a 15 km race, so this was surprising to me.

The coefficients in the regression model with predictors `Gender`, `Age` were not surprising: males and lower age were associated with faster times. The model in this case was approximately:

$$\texttt{Time} = -14.42 * \texttt{Male} + 0.35 * \texttt{Age} + 88.07,$$

where `Male` is the indicator 0-1 variable for males. There is possibly some selection bias from this process that renders the p-values meaningless, but they were all essentially 0.

4.2. **Predicting Split Differential.** As predictors for `Half_split_differential` using `Time` as a control, the lowest error subset out of `Gender`, `Age`, `Humidex`, was the full set {`Gender`, `Age`, `Humidex`}, and the smallest subset within one SE of the best was {`Age`}, with {`Gender`} very close behind.

We also considered all the subsets including or omitting `Time` and a constant variable `Constant`. Out of all the variables, `Time` was the most predictive variable, in the strong sense that every subset with it performed much better than every subset without it. After `Time` was included, a constant always helped. It was unclear to me that this would be the case—it seems somewhat reasonable that split should be purely proportional to time with adjustments for other variables.

The regression model with all the variables was approximately

$$\texttt{Half\_split\_differential} = 33.10*\texttt{M} - 1.62*\texttt{Age} + 0.78*\texttt{Humidex} + 6.06*\texttt{Time} - 381.71,$$

again with p-values all essentially 0.

A very interesting conundrum presents itself at this point: the coefficient for `M` is rather large and positive, indicating that for fixed values of the other variables, males tend to have more positive splits than females. As this seems to directly contradict the visualization above (which clearly indicated that males run less extreme positive splits than females, even after scaling for time), I did a little more digging to try to get a handle on the problem.

To simplify things a bit, I considered a regression with the only independent variables being `M` and `Time`. Again, the coefficient `M` was large and positive (about 28.8). Next, I tried a regression using only a constant variable and `M` to predict `Split_scaled`. This seemed to recover the observation from the visualization, since the `M` coefficient was negative (about -0.31), but on further investigation, it became clear that this is a pretty unreasonable model. (In general, it is usually ill-advised to use a ratio in a regression model—see the fantastic paper [Kronmal, Richard A., "Spurious Correlation and the Fallacy of the Ratio Standard Revisited" J. R. Statist. Soc. **156** (1993)].) Here, the underlying model is equivalent to

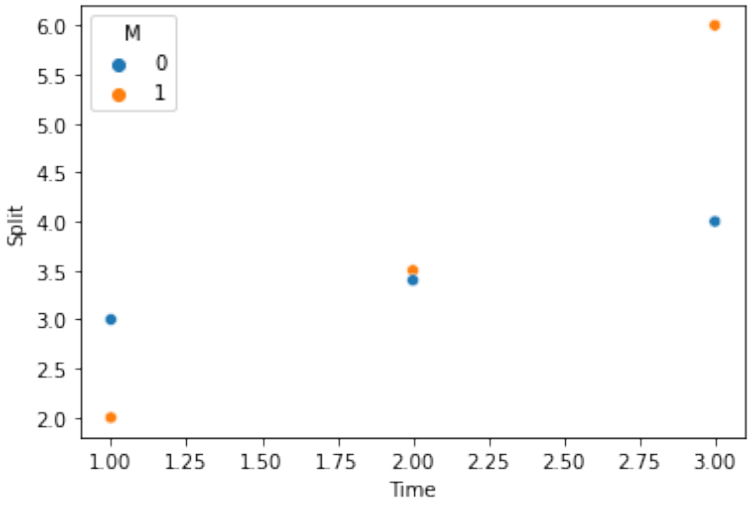
$$\texttt{Half\_split\_differential} = \beta_1(\texttt{M} * \texttt{Time}) + \beta_2(\texttt{Time}) + \texttt{Time} * \epsilon,$$

where ϵ is an independent error term with mean zero. In fact, running a regression using the same model as this with the last term replaced by ϵ alone still yielded a negative coefficient for $\texttt{M} * \texttt{Time}$. The interpretation here is that if you think of `Half_split_differential` as being purely proportional to `Time` (which is not that unreasonable of an idea), then that proportion is positive and a little bit less for males than females. This is interesting, but has the problem that we know from above that it doesn't lead to very good predictions if one assumes that `Half_split_differential` is purely proportional to `Time`: the cross-validation results are clear that a constant variable is important when predicting `Half_split_differential`. Also, the interaction variable $\texttt{M} * \texttt{Time}$ has a very high p-value (0.533) when included in a model with `M`, `Time`, and `Constant`, indicating it is not a useful predictor.

So who runs more well-paced races, males or females? The answer is subtle and depends on how you interpret this question. Loosely speaking:

- (Academic interpretation.) If you're interested in the pure splits (or the scaled version), then the answer is males. Males in general run less extreme positive splits, even after scaling by time.
- (Pragmatic interpretation.) If you're trying to predict splits, then the answer is females. If you take a male and female at a fixed time, it's a better guess on average that the male will run a more positive split.

A toy version of the situation is modeled as on the following page (the blue and orange dot in the middle can be taken to have the same value). The males (orange dots) have smaller scaled split values on average than the females (blue dots), but at a random fixed time, it is on average better to guess that the male will have a more positive split.


M
0
1
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
Split
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Time