20 November 2022

Georgia Institute of Technology

Fall 2022

MGT 6203 Group Project Final Report

INCREASING BOSTON MARATHON PROFITS THROUGH PREDICTIVE ANALYTICS FOR RACER FINISH TIMES

TEAM 5

BESHIR AISSI

JACOB BURDICK

DANIEL CIROCCO

ADAM GEORGE

QING LU

Contents

Background

The Boston Marathon is an annual 26.2-mile race conducted by the Boston Athletic Association (BAA). This prestigious event is typically held in mid-April in downtown Boston and attracts up to 30,000 participants each year from all over the world. Registration fees are typically $225-235 per participant. Figure 1 shows the route that racers follow during the event (Boston Athletic Association, 2022).

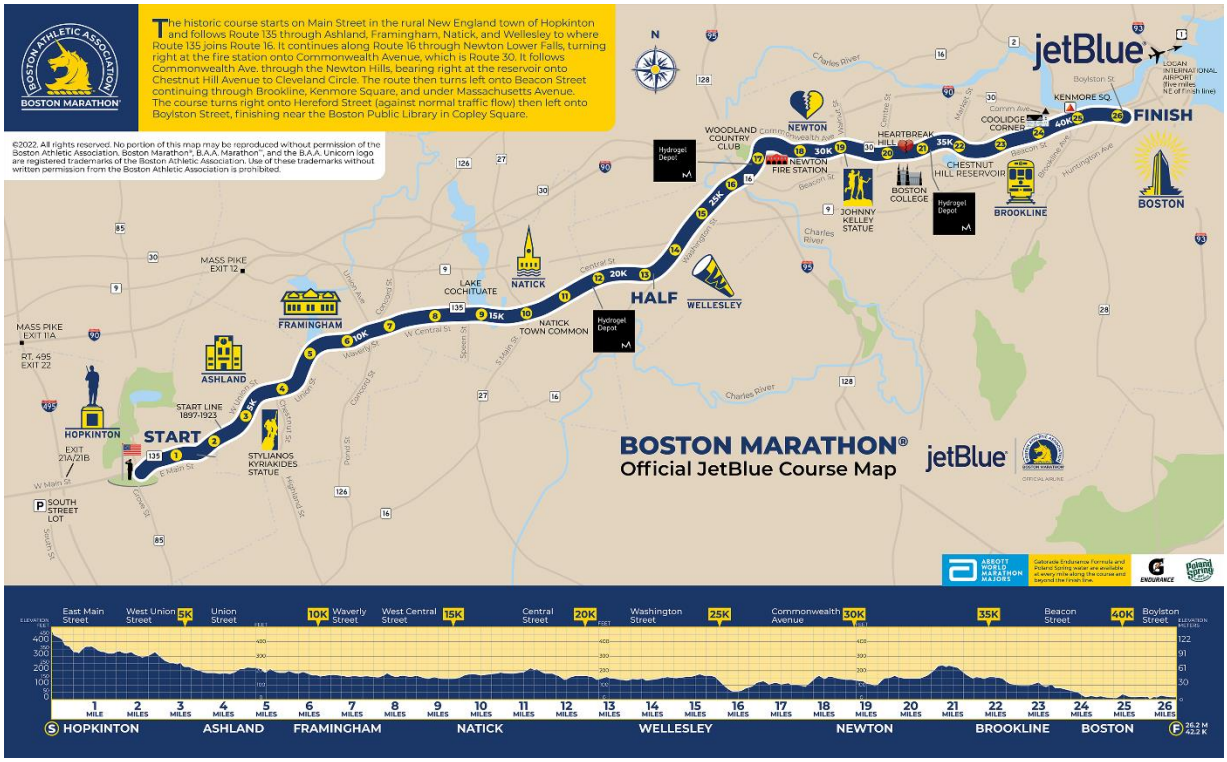

Figure 1: Boston Marathon route

Each year, it hands out hundreds of thousands of dollars in prize money and raises millions of dollars for charity. But hosting an event such as this is a massive undertaking—it requires coordination between the BAA and local government agencies to block off the route to vehicle traffic during race hours, arrange transportation and security, provide emergency medical services, and assemble sponsors and vendors. The organization must also setup and supervise staging zones, hydration stations, and recovery areas, as well as complete critical tasks such as registering participants, managing spectators, distributing medals to finishers, and liaising with security and medical personnel throughout the day. Completing these responsibilities requires a large staff dedicated to ensuring the race goes smoothly. In addition, small street sizes and strict time limits on when streets can be blocked off are logistical considerations that event organizers must manage each year as they set start times and deploy staff. Historically, participant completion times range anywhere from two to eight hours, with most participants finishing in 3-4 hours, as can be seen in Figure 2 below.

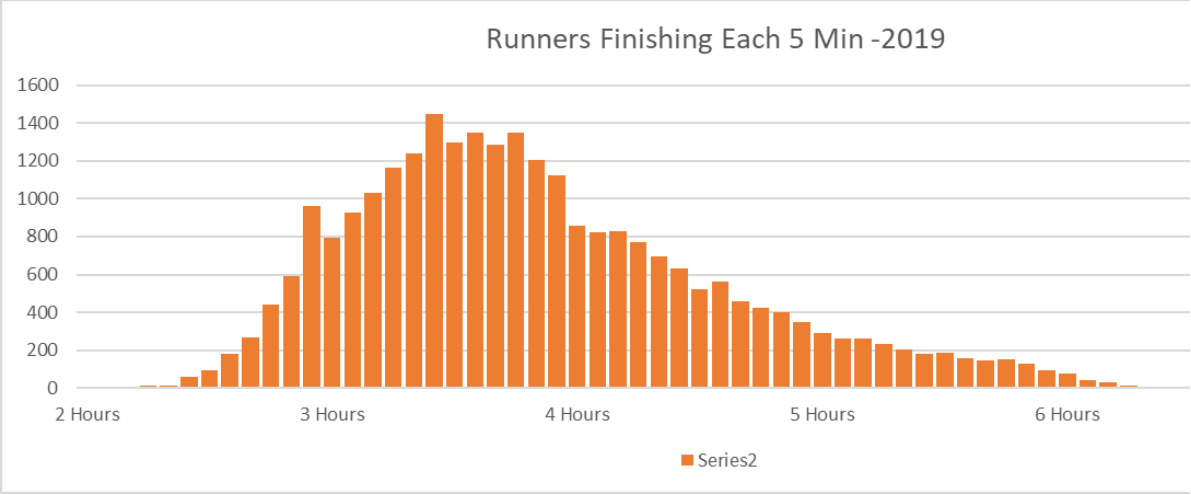

Figure 2: Histogram of number of finishers by 5 min increment in 2019

Objective

Our team seeks to provide recommendations to the BAA on ways to optimize various aspects of the Boston Marathon operations to maximize profits. Specifically, staffing is a large expense for any marathon, so reducing the paid staffing would reduce costs. To accomplish this, we will use several analytical models to predict racer finish times based on known characteristics of the participants and expected weather conditions. Because required staffing levels are tied to the maximum number of racers in an area at any given point in time, having accurate predicted finish times for race participants will allow the BAA to adjust logistical variables to minimize the number of staff necessary at race stations throughout the day.

Initial Hypotheses

We believed that our analysis would show that 'Age' and 'Gender' are the most significant factors in predicting a racer's finish time. This hypothesis was based on literature sources such as "Marathon Runners: How do they age?" as well as some basic reviews of the race data.

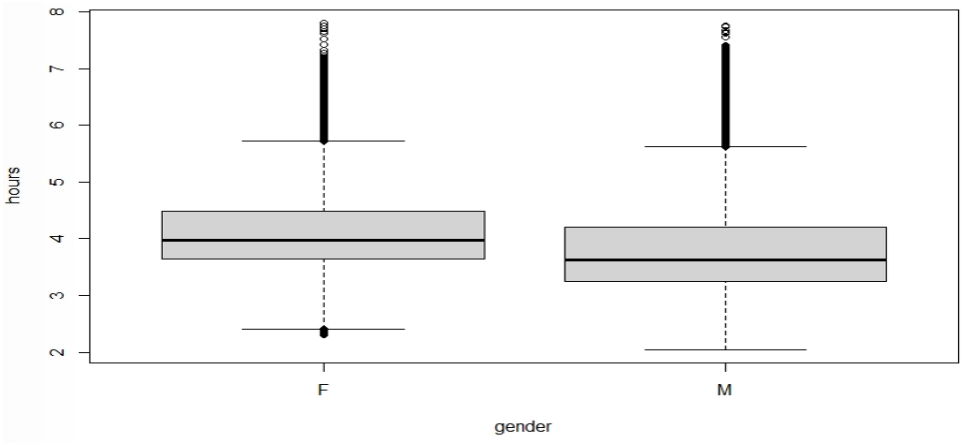

Figure 3: Boxplot of finish times by gender

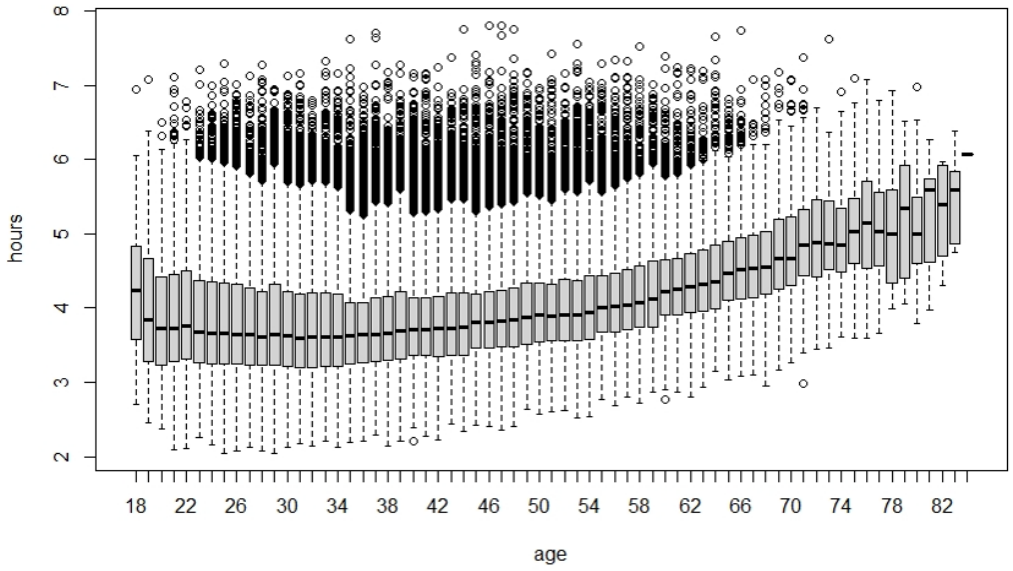

Figure 4: Boxplot of finish times by age

Figure 3 and Figure 4 show clear differences in finish times by gender and by age respectively. This matches our findings in literature and we believed these factors would prove to be significant in our models.

Additionally, we felt it was likely that 'Temperature' would have a notable impact on performance as well. This was based on the article "Impact of weather on marathon-running performance", which found that as temperature increased from 5 to 25 deg C, racer performance progressively decreased (MR Ely, 2007).

Methodology

Data Sources

Adrian Hanft has collected Boston Marathon results and participant demographics going back to 1897 in a public GitHub repository (Hanft, 2019). Our team used data from years 2010 through 2019—a dataset of over 260,000 individuals—to develop our predictive models. The dataset includes age, gender, and city of residence information for each participant, as well as their race completion time.

Additionally, the website *WUnderground* provides detailed hourly historical weather data. We collected a variety of weather parameters for the same ten-year timeframe to include in our models.

Finally, a public GitHub repository of the top 1000 cities in the United States (based on population) was a third dataset that was incorporated into our analysis. We used these data to help us determine if participants resided in the US or abroad.

Data Cleaning

Our team began by collecting and cleaning the race data. Using R, the data for each year were loaded into a data frame and formatted to a standard structure by adding a column for 'Year' and renaming

columns as necessary due to variations year to year in the naming conventions used in the data files. During this process, we noticed that two different files were available for year 2013 - one with and without 'diverted' results. Further research into 2013 quickly identified this as the year of the Boston Marathon bombing, which caused many racers to be diverted away from the finish line (CNN, 2013). Given these abnormal events, we decided to treat this entire year an outlier and exclude its data from the analysis.

Once data from all years was properly formatted, all data was appended to a single data frame. The dataset was further prepared by dropping several rows where 'Gender' was not 'M' or 'F' (indicating an issue with parsing the data file for that row), setting data types for the columns, and creating a binary dummy variable 'Male' that has a value of 0 or 1. A Rosner test was done on the data to identify a handful of outlier points, which were removed from the dataset.

To transform the participant's city of residence information into a useful model factor, each location was checked against the dataset of Top 1000 US Cities. A new factor, 'Is Top US', was created and set to TRUE when a match (case-insensitive) was found, and FALSE when no match was found. While not accurate for 100% of the participants, this factor was used to approximate when a runner was US-based or international. The city factor was then dropped from the dataset.

To prep the weather data, data for all years were combined to a single dataset using python. Then, the data were filtered to only include data between 8 AM and 5 PM (the hours during which the race is run). Several columns had to be formatted by removing units from the value and setting the data type for the column to numeric. Additionally, any 3-letter designations for 'Wind Direction' were simplified to the nearest cardinal direction (e.g. SSW to S). Finally, the data for each year were aggregated to a single value for each factor. Numeric factors such as 'Temperature' and 'Wind Speed' were averaged over the time period, and the mode was taken for categorical factors such as 'Wind Direction'.

A correlation matrix was created to identify any variables that were highly related. As can be seen in Figure 5, 'Wind Gust', 'Humidity', and 'Precipitation' all correlated strongly with 'Wind Speed', so we elected to remove those three variables from the analysis.

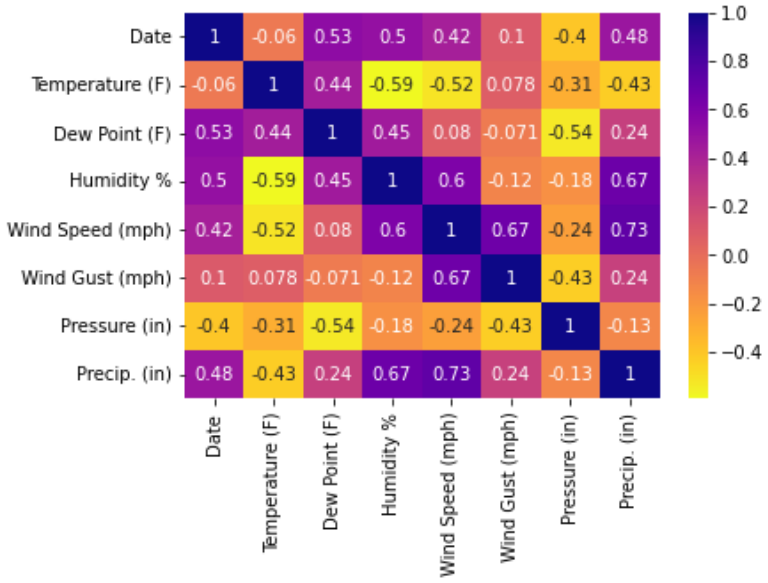

Figure 5: Correlation matrix of weather variables

Lastly, using R, the race results data and weather data were joined together on date to provide a consolidated dataset, and categorical variables were cast as factors to prep them for our analysis.

Feature Selection and Model Creation

Based on the research our team conducted into factors that impact race performance, we initially selected 'Age', 'Gender', 'Temperature', 'Wind Direction', and 'Weather Condition' to be included in our model. This research also informed our decision to group temperature into low (42.9-54.3), medium (54.4-65.6), and high (65.7-76.9) buckets and treat it as a categorical variable rather than a continuous variable. Additionally, we decided to include the 'Is Top US' variable in our analysis, even though we expected it would not be significant and would end up being dropped during feature selection.

To identify which features were significant to our model, a forward stepwise regression, backward stepwise regression, and combined stepwise regression were run using R. These regressions provided AIC values for each factor based on its importance. Guided by those AIC values, linear regression models were generated by adding in the six factors one at a time in order of importance. When comparing the 6 models based on R-squared value, we found that the model that included all 6 factors performed best. Additionally, the p-values for that model indicated that all the factors were significant at a greater than 99.9% confidence interval, so no features were dropped from the final model.

Model Interpretation

The best performing model is shown below in Figure 6. It included all 6 features, and the R-squared value was 0.1607.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.8547 -0.4638 -0.1771  0.2989  4.0997

Coefficients:
                                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      3.2880603  0.0070788  464.494   <2e-16 ***
age                              0.0179219  0.0001299  137.947   <2e-16 ***
genderM                         -0.4296565  0.0029174 -147.274   <2e-16 ***
temperature(54.3,65.6]          -0.0615706  0.0050751  -12.132   <2e-16 ***
temperature(65.6,76.9]           0.2357000  0.0084969   27.740   <2e-16 ***
wind_directionNS                 0.0989073  0.0089912   11.000   <2e-16 ***
wind_directionW                 -0.0506661  0.0060911   -8.318   <2e-16 ***
weather_conditionLight Rain / Windy  0.1839059  0.0059462   30.929   <2e-16 ***
weather_conditionMostly Cloudy   0.1356565  0.0086792   15.630   <2e-16 ***
weather_conditionPartly Cloudy   0.1956743  0.0050738   38.566   <2e-16 ***
in_top_usTRUE                   -0.0760052  0.0030531  -24.895   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6817 on 231661 degrees of freedom
Multiple R-squared:  0.1607,    Adjusted R-squared:  0.1607
F-statistic:  4435 on 10 and 231661 DF,  p-value: < 2.2e-16
```

Figure 6: Model summary

While this value seems on the low side, it is difficult to determine what a "good" R-squared value would be for this scenario. As Figure 7 demonstrates, even within a single Age and Gender group for one year (so consistent weather conditions), there are a large number of runners and a broad distribution of finish times.

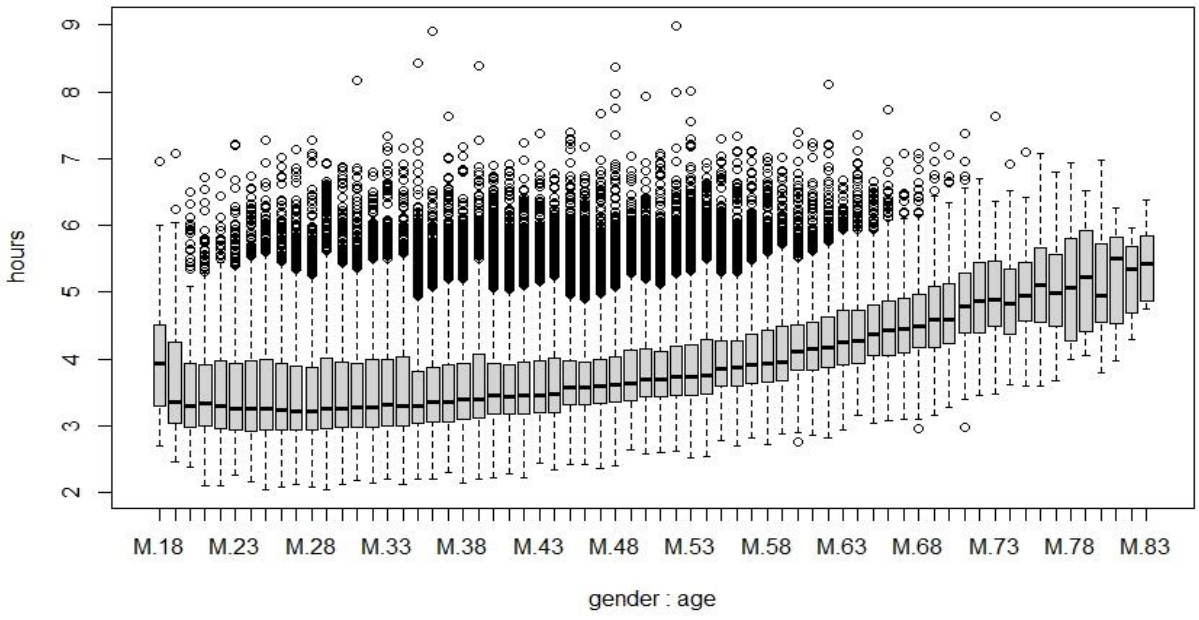

Figure 7: Boxplot of male racer finish times by age

Considering this, there are clearly other factors that impact finish time that are not comprehended by our model, so the low R-squared value, which is a measure of how much of the data's variance is explained by the model, makes sense.

To review the error in our model, we compared our predicted finish time for each runner against their actual finish time. The resulting histogram can be seen in Figure 8.

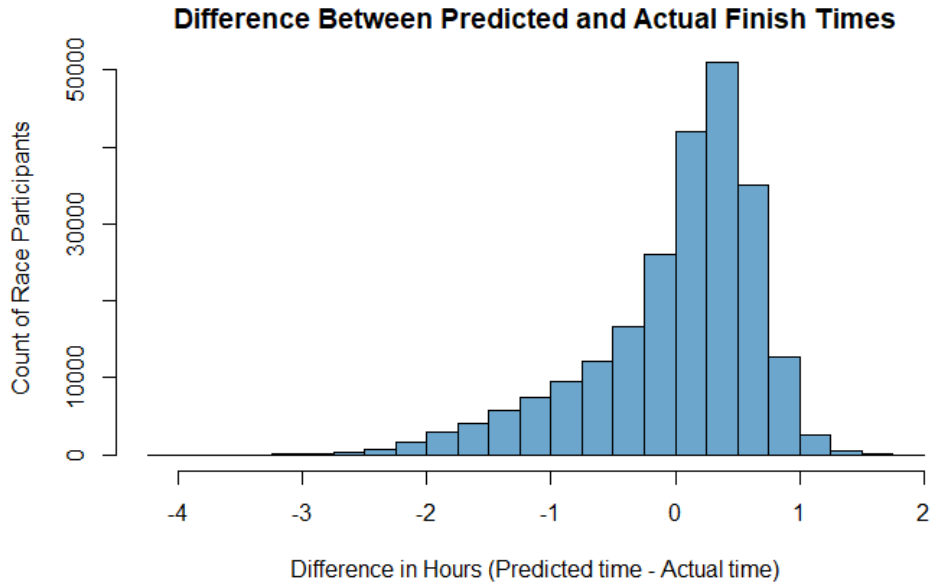

Figure 8: Histogram of error in predicted finish times

To put this in numerical terms, our model predicted participant finish times within 7.5 mins of their actual finish times for only 46% of the participants. If we allowed for up to 30 mins of error, the model correctly generates predictions for 78% of participants.

Just because our model has a low R-squared value does not mean it is not useful, however. When we create a throughput curve from our predicted finish times, we see that it still follows the same distribution seen back in Figure 2, which shows that almost 60% of runners finish the race in 3-4 hours. Even allowing for 30 mins of error, we can confidently identify 30% of the overall participant population whom we expect to finish in the middle of this time window. Using this information, the start times for those participants can be staggered throughout the race throughput rate more consistent.

Additional Findings

Reviewing the coefficients for our model, shown back in Figure 6, provided some additional insights into the factors which most affect performance. These are summarized below in Table 1.

Factor	Effect
Gender (base case = Female)	Male -25 min
Temperature (base case = Low)	High +13 min Medium -4 min
Weather Condition (base case = Cloudy)	Rainy/Windy +11 min Partly Cloudy +11 min Mostly Cloudy +8 min
Wind Direction (base case = E)	N/S +5 min W -3 min
Resides in US (base case = False)	True -4 min
Age	+1 min

Table 1: Model factor effects

It is clear that 'Gender' was the most significant factor, as male runners finish 25 mins faster than female runners. This was not unexpected and aligns with our initial hypothesis and research. Additionally, we see that 'Temperature' has a large effect performance, which again matches our hypothesis. However, the effects is not linear - high temperatures negatively impact performance the most, but low temperatures have a negative effect as well. Moderate temperatures are best, and this matches our findings in the literature.

'Weather Condition' turned out to be more impactful than expected. While Rainy/Windy conditions contributing to slower race times intuitively makes sense, seeing large negative impacts from Partly Cloudy and Mostly Cloudy conditions (vs the base case of Cloudy) was curious. Our hypothesis is that less cloudy conditions contribute to higher temperatures, which has already been shown to negatively impact performance.

'Age' was the final variable we hypothesized to be significant, and it proved to be an interesting factor. Our model predicts that every year accounts for 1 added minute of race time. However, looking back to Figure 7, there is little to no drop off in performance related to age between approximately ages 20 to 40, at which point performance starts to decay. This generally supports the assertions made in article "Marathon Runners: How do they age?" referenced earlier which claimed peak performance between age 25-35 (Trappe, 2007), as well as the article "Age-related Changes in Marathon and Half-Marathon Performances", which proposed a slightly wider band of age 20-49 where performance is roughly equivalent (D. Leyk, 2007). In hindsight, perhaps model performance could have been improved by splitting our data based on age Age <=40 and Age > 40, then generating models for these groups independently.

Finally, the 'Is Top US' (Resides in US) variable yielded the most surprising result. We initially did not believe this factor would be significant at all, but our model predicts that runners who live in the US finish 4 minutes faster than international runners. We do know that this result is questionable because

our joining of 'City of Residence' information against the list of top 1000 US cities is really only a best guess approximation of country of residence. For one, there are many city names that can exist in both the US and foreign countries. For another, there are still tens of thousands of US cities unaccounted for that could cause a US participant to be misclassified as international. However, it is possible that this effect is real and is worth exploring in more detail. A possible hypothesis is that US-based runners have an easier time traveling domestically to the marathon, finding accommodations, foods they like, and interacting with others, and this allows for a less stressful leadup to the race, resulting in better performance on race day.

Figure 9 below consolidates of 'Temperature', 'Gender' and 'Resides in US' to show the combined effects of these factors.

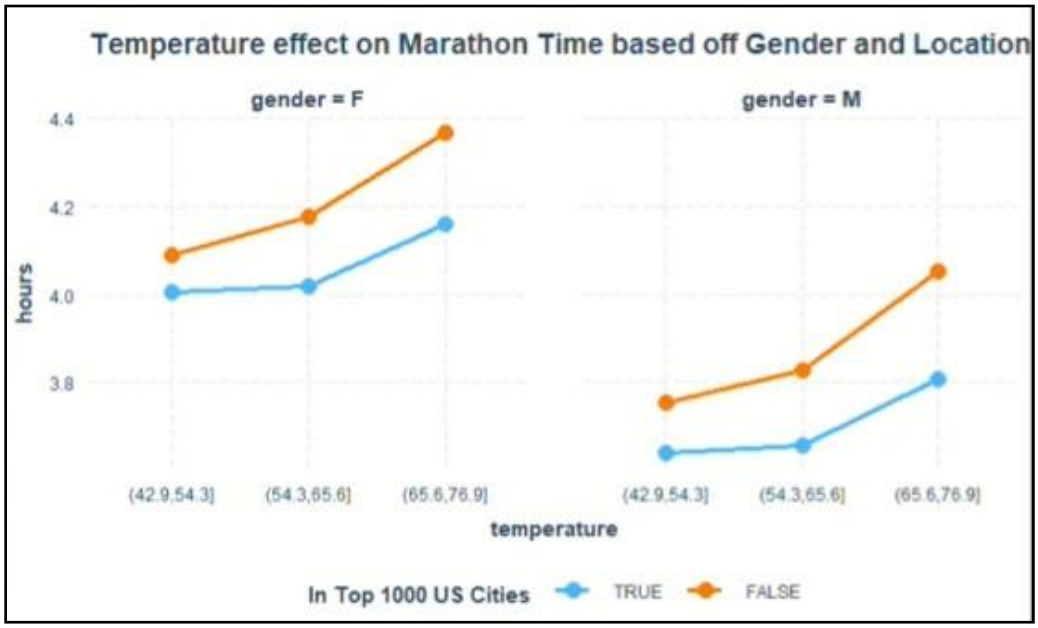

Figure 9: Trend chart of race time by gender, temperature, and city of residence

Future Research

To further improve the predicted finish times of runners, it would be valuable to have additional personal information for each of the participants. Factors such as ethnicity, income (indicating their ability to spend money on race preparation), prior marathon experience, prior results, amount of training done, diet, injury history, training elevation, distance traveled, and even less obvious items such as shoe brand and model could all provide key insights into racer performance. These data could all be collected during signup or at pre-race registration with minimal effort.

Additionally, having better visibility into the operations of the Boston Marathon would be useful. More accurate figures around profit margins for the race, typical numbers of staff and their distributions on race day, detailed task lists, and payroll information could help our team provide more detailed and directed recommendations on ways for the BAA to increase profits.

Final Conclusions

Through our analysis, we developed a regression model that can effectively predict a participant's finish time within 30 minutes with 80% accuracy. While not highly useful when considering a single runner, it is suitable for modeling the distribution for the entire population of runners. Using this information, we can select a subset of runners who are expected to finish in the same high-volume timeframe and spread them out through the entire race day. This will reduce the maximum throughput rate of racers throughout the day. Lower throughput rate correlates to lower staffing needs.

We expect that by providing a more consistent throughput rate, we can reduce the staffing requirements at the finish line by 50%. Assuming that currently 200 staff are required for 4 hrs at $16/hr, we estimate that about $6400 could be saved at the finish line area alone. Figure 10 below illustrates this concept.

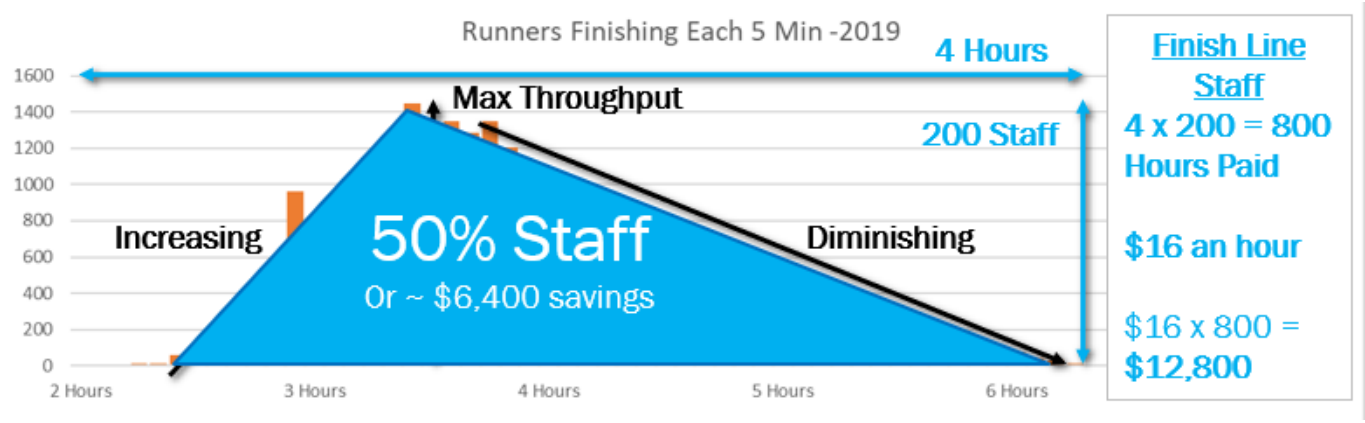

Figure 10: Savings due to staff reductions

If we expanding these assumptions to the entire race course, there are likely to be another 200 staff handling tasks at the starting line and stations along the race course. If these stations don't need to accommodate as high of a throughput rate of runners, they can also reduce their staff by 50%. Overall, our team believes we can save the BAA approximately $13,000 in staffing costs.

Works Cited

Boston Athletic Association. (2019, April 19). *2019 Boston Marathon Injects More Than $200 million into Greater Boston Economy*. Retrieved from Boston Athletic Association Web site: https://www.baa.org/2019-boston-marathon-injects-more-200-million-greater-boston-economy

Boston Athletic Association. (2022). *Boston Marathon Course Information*. Retrieved from Boston Athletic Association Web site: https://www.baa.org/races/boston-marathon/enter/course-information

CNN. (2013, April 22). *Boston Marathon terror attack*. Retrieved from CNN Web site: http://www.cnn.com/interactive/2013/04/us/boston-marathon-terror-attack/

D. Leyk, O. E. (2007). Age-related Changes in Marathon and Half-Marathon Performances. *International Journal of Sports Medicine*(28), 513-517. doi:10.1055/s-2006-924658

MR Ely, S. C. (2007). Impact of weather on marathon-running performance. *Medicine and Science in Sports and Exercise*(39), 487-493. doi:https://doi.org/10.1249/mss.0b013e31802d3aba

Page, V. (2022, May 12). *The Economics Behind Marathons*. Retrieved from Investopedia: https://www.investopedia.com/articles/investing/100815/economics-behind-marathons.asp

Trappe, S. (2007). Marathon Runners: How Do They Age. *Sports Med*(37), 302-305. doi:https://doi.org/10.2165/00007256-200737040-00008

Data Sources

Boston Marathon Historical Race Results Raw Data

Hanft, A. (2019, October 23). *Boston Marathon Data Project.* Retrieved from GitHub: https://github.com/adrian3/Boston-Marathon-Data-Project

Weather Historical Raw Data

WUnderground. (various years). *East Boston, MA Weather History*. Retrieved from WUnderground: https://www.wunderground.com/history/daily/us/ma/east-boston

Top 1000 US Cities Raw Data

https://gist.github.com/Miserlou/11500b2345d3fe850c92

Team 5 Code

https://github.gatech.edu/MGT-6203-Fall-2022-Canvas/Team-5/tree/main/Final%20Code

Team 5 Cleaned Data

https://github.gatech.edu/MGT-6203-Fall-2022-Canvas/Team-5/tree/main/Data

Team 5 Final Presentation

https://github.gatech.edu/MGT-6203-Fall-2022-Canvas/Team-5/tree/main/Final%20Presentation%20Slides