David Schneider

DA 516 Introduction to Machine Learning

December 9, 2020

**The Impact of Runners’ Age and Gender on Boston Marathon Finishing Times: Exploring the 2019 Boston Marathon with Descriptive and Inferential Statistics and Machine Learning**



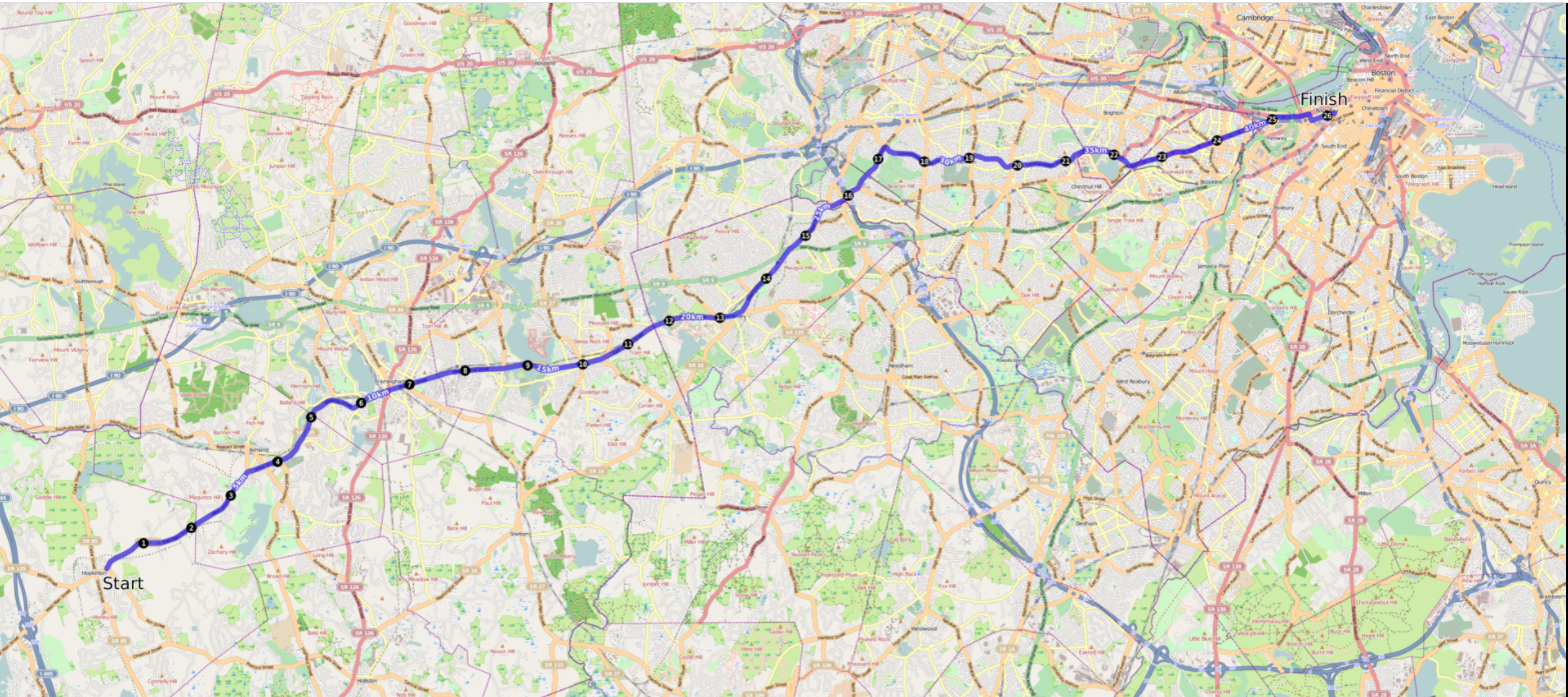
*Photo Credit: WCBV TV,* <https://www.wcvb.com/article/boston-coronavirus-response-update-mayor-walsh-may-28-2020/32699908>

**Abstract**

This project analyzes data on the 2019 Boston Marathon finishers using descriptive and inferential statistics, linear regression, and binary classification algorithms. It explores the extent to which a runners’ finishing time changes with age and determines that finishing time increases by 5 minutes with each year of age. The Mean Squared Error of a linear regression is 0.025, which can be improved to 0.0002 when using a decision tree regressor. This project also classified finishers into men and women based on the runners’ age, finishing times, and other features. All classification algorithms performed better than the default accuracy measure of .55 for men and .45 for women, with the random forest classifier correctly classifying gender with .995 accuracy rate. However, readers should take care not to over-generalize on the dataset which contains a mix of elite runners and non-elite runners raising money for charity and represents only one year’s worth of marathon results.

**Background**

The Boston Marathon is the world's oldest annual marathon and ranks as one of the world's best-known road racing events**.**  It was established in 1897 and is held on the third Monday in April. The 2019 results are the most recent results available as the event was cancelled in 2020 due to the COVID-19 pandemic (the 2021 running has also been postponed to the fall of 2021). In the past ten years, the marathon has attracted between 26,000 and 30,000 entrants. The race begins in Hopkinton MA and finishes at Copley Square. Here is a map of the course:



Qualifying for the Boston Marathon is highly competitive. The Marathon establishes qualifying times based on age and gender and runners must demonstrate they have achieved a qualifying time in a marathon run in the past two years. The qualifying times for the 2019 Boston Marathon are shown in the charts below:

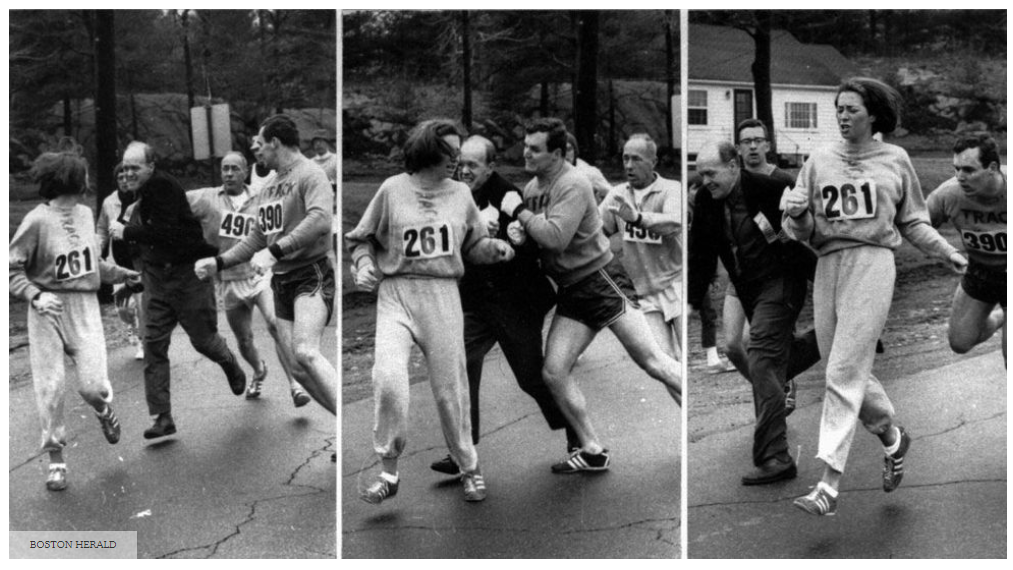


*Source: The Boston Athletic Association,* [www.baa.org](http://www.baa.org)

In addition to runners who qualify based on performance, about one-fifth of the marathon's spots are reserved each year for charities, sponsors, vendors, licensees, consultants, municipal officials, local running clubs, and marketers (referred to in the rest of this paper as “charity runners”). In 2020, Charity Runners were required to raise at least $5,000 from charities designated by the Boston Athletic Association. (BAA)

Up until the 1970s, men, considered women too physiologically frail to run marathons, but that did not stop some women from competing. Roberta Gibb was the first woman to run the full Boston Marathon in 1966. Gibb, who did not run with an official race number during any of the three years (1966-68) that she was the first female finisher, hid in the bushes near the start until the race began.

In 1967, Katherine Switzer did not clearly identify herself as a female on the race application and was issued a bib number. As shown in the photo below, BAA athletic director Jock Semple tried to physically remove Switzer from the race once she was identified as a female entrant. Switzer’s boyfriend, tackled Semple and Switzer finished the race (she and Semple eventually reconciled).



*Photo credit: the Boston Herald*

When the Amateur Athletic Union (AAU). Permitted marathons (including Boston) to allow women entry in the fall of 1971, Nina Kuscsik’s 1972 B.A.A. victory the following spring made her the first official champion. Eight women started that race and all eight finished. Women’s participation has steadily increased to the point where women comprise around 45% of the entrants in all marathons since 2014, according to a survey by RunnerClick, which looked at data on marathons run between 2014 and 2017.

Over time, the Boston Marathon field has also become more diverse by age, geography, national origin, and ability. Marathon runners now regularly include entrants in their 70s and 80s and include runners from throughout the United States and the world. The marathon also established wheelchair division competitions and divisions for handcyclists and runners with visual impairments and mobility impairments.

The weather and temperature of the Boston Marathon can vary dramatically and impacts the performance of the field. For the 2019 marathon temperatures ranged from the 50s-60s with rain at the beginning of the race.

**Research Questions**

This project examines the relationship between demographic features and finishing times and asks the following questions:

1) To what extent do runners finishing times change as they age?;

2) Is it possible to accurately identify a runner’s gender based on his or her age and finishing time?;

3) Which country, city, and state have the most runners?

4) Which country has the fastest runners?

Question #1 has practical implications insofar as it may help runners predict how much slower they may become as they age (and therefore how much harder they may need to train to qualify for the race). Questions #3 and #4 may help runners willing to travel in order to improve their performance identify cities, states, and countries with active running communities and running culture. Question #2 has little practical value, however it helped this aspiring data science get some practice with binary classification algorithms.

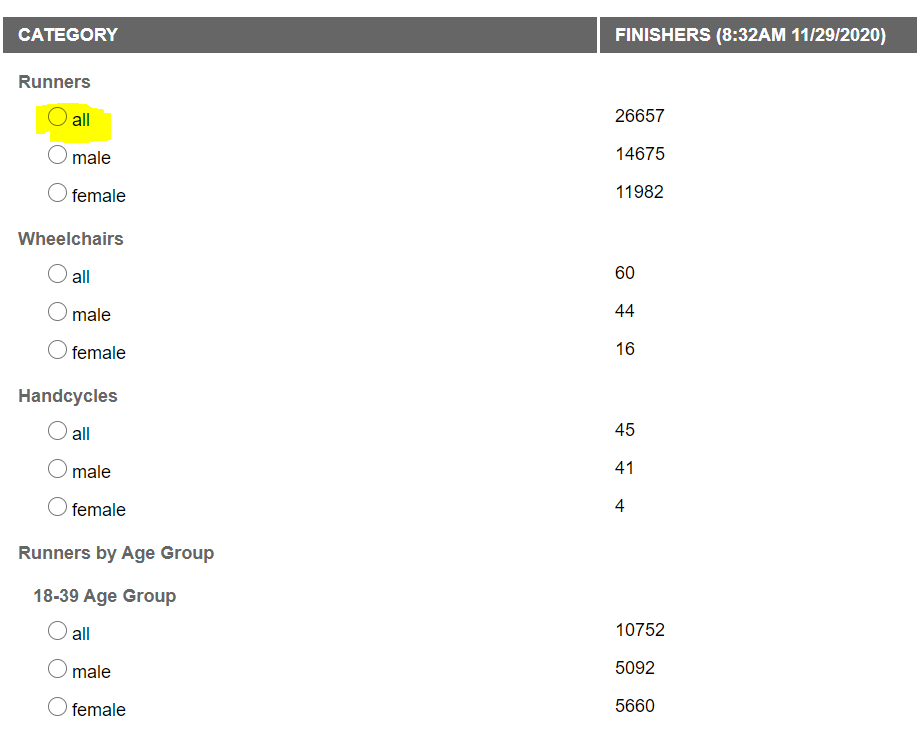
***Note: In the remaining sections, I’ve added a reference to the cell of my Jupyter notebook in italics and parentheses so that readers can reference my results back to the code I used.***

**Data Acquisition and Discovery**

I downloaded the data from the Boston Athletic Association website,

http://registration.baa.org/2019/cf/Media/iframe\_ResultsSearch.cfm?mode=entry

The website gives users the option to download data by gender, age, division, etc. I downloaded the file corresponding to all runners (see screenshot below).



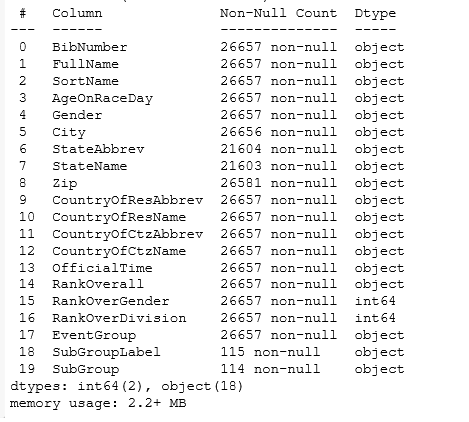
The data set includes 26657 records and 20 features. *(Cell 6).* A screen shot of the feature headings and first nine records *(Cell 5)* is below:



Although the BAA website does not provide a data dictionary or other metadata, I’ll provide a definition of the most relevant features in the data frame

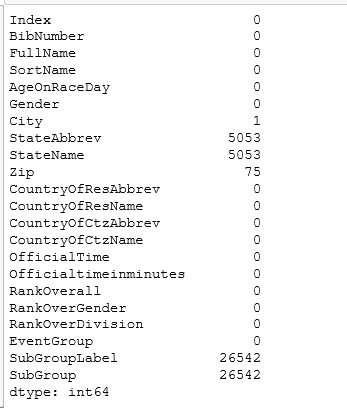
* Bibnumber: The number of the bib assigned to each runner. Bib numbers are a rough proxy for a runner’s ability as they are assigned on the basis of a runner’s marathon qualifying time. Runners with the fastest qualifying times are assigned to waves at the front of the pack.[[1]](#footnote-1)
* FullName: The first name and last name of the runner.
* AgeOnRaceDay: The runner’s age on April 19, 2019
* Gender: The runner’s gender (Male or Female)
* City: The city that the runner resides in
* StateAbbrev/StateName: The state abbreviation and full name of the state the runner resides in (includes US States and Canadian Provinces)
* Zip: The zip code corresponding to the runner’s address.
* CountryofResAbbrev/Countryof ResName: The abbreviation and full name of the country that the runner resides in.
* CountryOfCtzAbbrev/CountryOfCitzName: The abbreviation and full name of the country of which the runner is a citizen.
* Officialtime: The runner’s official finishing time.
* RankOverall: The place that the runner finished in relative to all other finishers.
* RankOverGender: The place that the runner finished in relative to other finishers of the same gender.
* RankOverDivision: The place that the runner finished relative to other members of his/her age division
* Subgroup label/subgroup: A category identifying subgroups of runners: Visually impaired, mobility impaired, and duos (a duo team is an able-bodied runner pushing a non-ambulatory person).

The data is a mix of numerical, categorical, and ordinal data, although when first downloaded, only two of the features are expressed as integers, with the rest expressed as objects *(Cell 9)*:



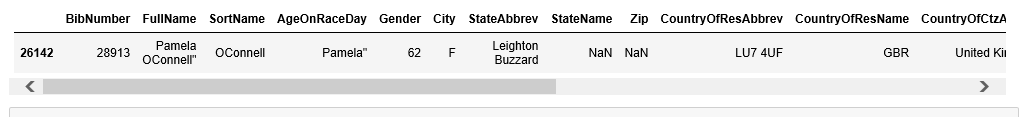
**Anomaly Detection and Data Cleaning**

The dataset is relatively tidy and a search for duplicate records revealed no duplicates *(Cell 10).* Some features are missing data, as shown in the table below *(Cell 29).*



The features with missing values are not essential to performing the machine learning analysis and were dropped as part of preparing the dataset for machine learning.

I identified an anomaly with one row in the data set (row 26142) where a mistake caused a runner’s age, gender, and other information to appear in the wrong order *(Cell 17):*

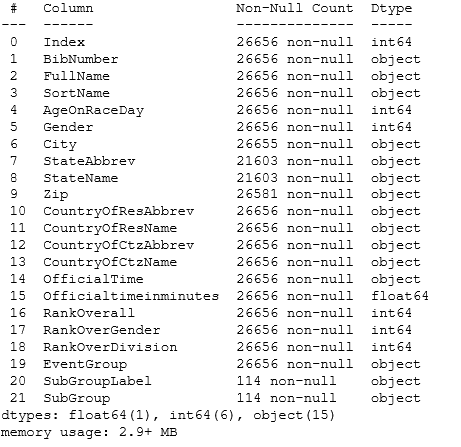


I probably could have found a way to align the data to the correct features but instead took the path of least resistance and dropped this record from the dataset *(Cell 18).*

**Feature Transformation**

I transformed the Age feature from an object to an integer *(Cell 19)* and replaced “M” and “F” in the Gender feature with “1” and “0” *(Cell 18).* I converted the Officialtime into a datetime Dtype *(Cell 20)* and then converted this number into minutes which I labeled Officialtimeinminutes *(Cell 21).*

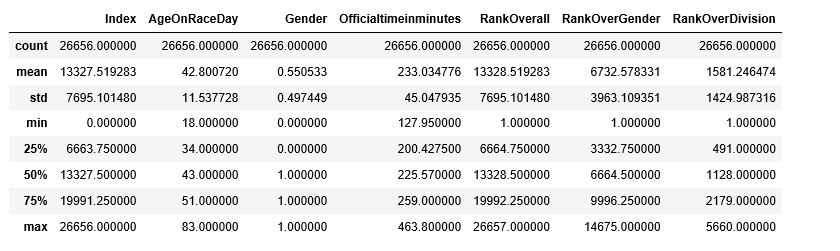
After these transformations had been completed, the data set consisted of the following Dtypes *(Cell 24):*



**Exploratory Data Analysis and Data Visualization**

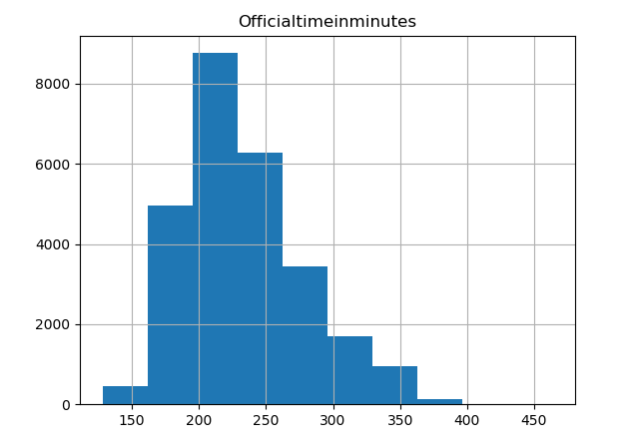
Summary Statistics

I ran a summary of the numeric data *(Cell 32)* to identify the mean finishing time and other summary statistics:



I was surprised that the mean finishing time as 233 minutes (or 3 hours and 59 minutes) since this is slower than the qualifying times of many of the runners. However, it likely reflects the participation of charity runners who made up around 20% of the field and whose finishing times are typically slower.[[2]](#footnote-2)

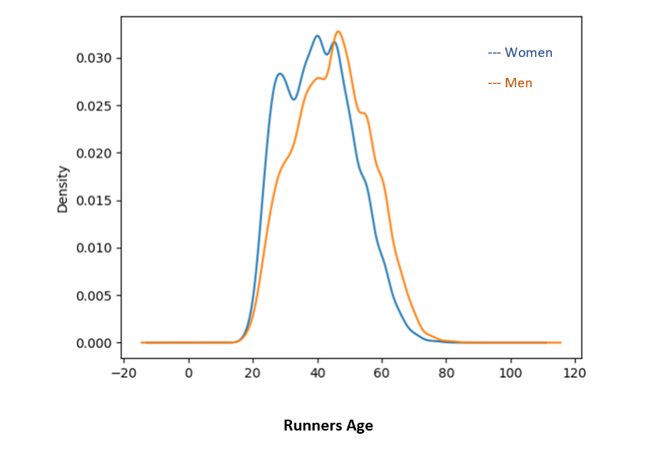
A histogram of finishing times for all runners *(Cell 34)* shows a roughly Gaussian distribution with a rightward skew:



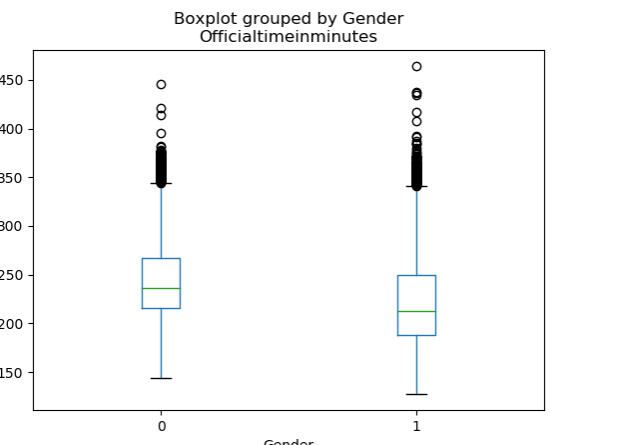
Age, Gender, and Finishing Time

The 2019 Boston Marathon finishers included 14,675 men and 11,981 women *(Cell 39)* for a split of 55% men and 45% women *(Cell 40).* This split is important because it is the default split for a binary classification based on gender. (In other words, if a classifier only guessed that a finisher is a woman, it would be right 45% of the time and if it only guessed an entrant is a man it would be right 55% of the time).

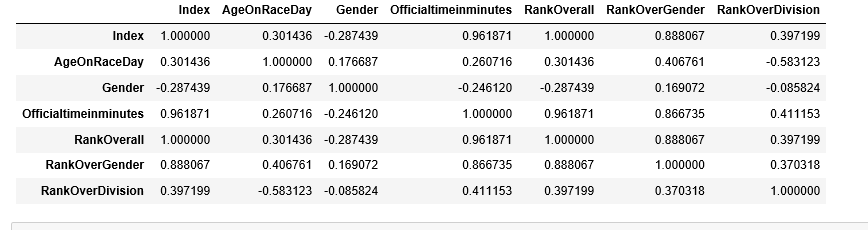
The mean age of men finishers is 45 years old with a standard deviation of 11 years *(Cell 44, 45)* and the mean age of women finishers is 40 years old with a standard deviation of 10 years *(Cell 44, 45).* The chart below plots the age distribution of men and women finishers *(Cell 35)*:



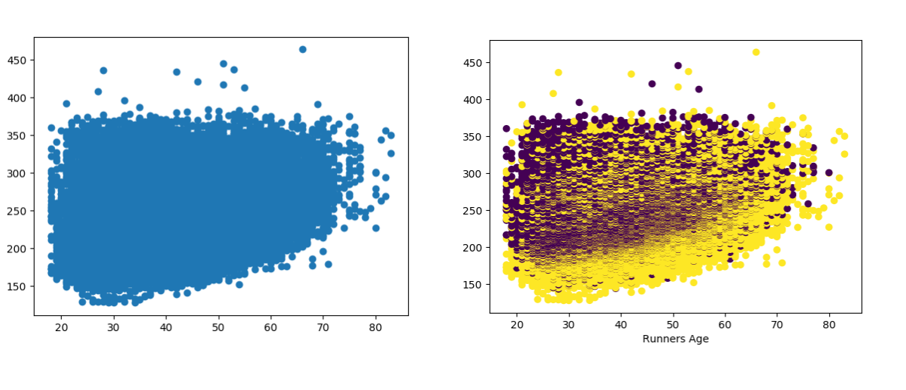
The mean finishing time for men is 223 minutes (3:49) with a standard deviation of 45 minutes and the mean finishing time for women is 245 minutes (4:12) with a standard deviation of 40 minutes *(Cell 41, 43)*. The median finishing time for men is 212 minutes (3:38) and median finishing time for women is 236 minutes (4:02) *(Cell 42)*. The median may be a better indicator of centrality since it reduces the impact of outliers. The boxplot below *(Cell 38)* summarizes finishing time by gender:



Correlations

The summary correlations for the numeric data *(Cell 33)* is shown below: 

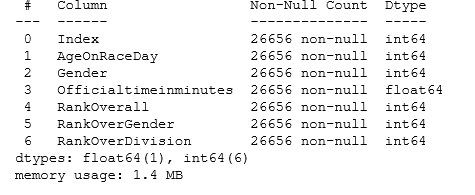
Not surprisingly, finishing time is highly correlated with rank. (the faster you run, the higher your finishing place). Age is positively but modestly correlated with finishing time and gender is modestly and negatively correlated with finishing time. The scatter plot on the left *(Cell 37)* shows the correlations between age and finishing time for all runners and the one on the right *(Cell 36)* shows finishing time by age and gender (with men shown in yellow and women in purple):



These data visualization don’t make it appear that gender, age and finishing time are particularly severable.

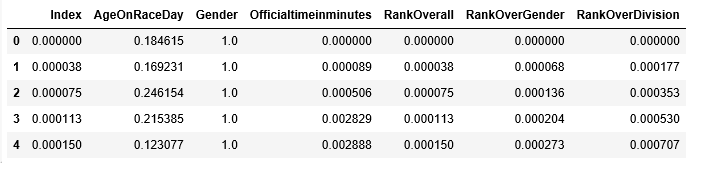
**Preparation for Machine Learning**

I dropped all of the non-numeric features, leaving only those features that were integers or floats *(Cell 794)*:



It seemed like a shame not to further analyze the categorical data, so I brought that information back into the data set for Part 5 of this project (“Going the Extra Mile: Additional Exploration of City, State, and Country and Finishing Time”).[[3]](#footnote-3)

I then scaled the features *(Cell 798)* and added labels *(Cell 800)* to produce the resulting data set:



I created a training set and a testing set using an 80/20 train/test split *(Cell 843)*. This resulted in a training set with 21,324 instances and a test set with 5,332 instances *(Cell 844, 845).*

Finally, I modified the training and testing sets by dropping the dependent variable, “Officialtimeiniminutes” from the training set and creating a version of the training set with only this variable *(Cell 885).* I also took these steps with the testing set (Cell 925).

**Algorithm Selection**

Regression

I started my regression analysis with a simple OLS linear regression using the statsmodels package *(Cell 849)* with AgeOnRaceDay as the independent variable and Officialtimeinminutes as the dependent variable *(Cell 851).* This produced the results shown below, with a coefficient of 5.1448 minutes (meaning finishing time increases by 5.1 minutes with each additional year of a runners’ age) and 95% confidence that the true coefficient is between 5.127 and 5.163. The p-value of the regression is 0.00 and the R-squared is .923 meaning that around 92% of the variance in the finishing times can be explained by this model.



I ran similar linear regressions with data sets comprised of only men runners *(Cell 856)* and only women runners *(Cell 859)* and produced similar results. The coefficients for the men-only data is 4.774 *(Cell 857)*  and for the women only data is 5.69. *(Cell 860).*

The results showing runners getting 5 minutes slower for each year older initially surprised me since it would mean that most runners would fail to re-qualify for the next year’s Boston Marathon based on the qualifying standards set for the BAA. However, (as a reminder) the dataset includes the charity runners who may have a wider range of finishing times than the qualifying runners.

I calculated a mean squared error of 67.7255 and a mean absolute error of 52.5217 for the linear regression, results that I tried to improve upon in the next section of the report (Fine Tuning Algorithms).

Binary Classification

I began my binary classification analysis by importing a decision tree classifier *(Cell 414)* and selecting all of the feature columns as the features and gender as the target variable *(Cell 415)*. This produced an accuracy of 0.987 *(Cell 417).*

**Fine Tuning Algorithms**

For the regression analysis, I sought to reduce the mean squared error by selecting additional features and using different regression algorithms. I imported the LinearRegression model *(Cell 863)* and ran the linear regression using all of the features in the data set against the Officialtimeinminutes dependent variable. *(Cell 864).* This resulted in a MSE of 0.0361 and an absolute error of 0.0257.

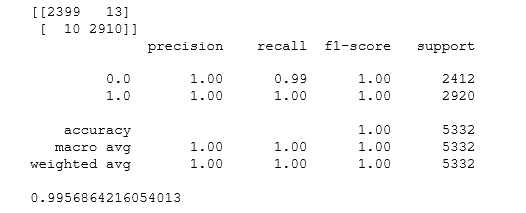
I then imported a decision tree regressor *(Cell 900)* which produced a mean squared error of 0.0002.

Here is a summary of the performance of the regression models:

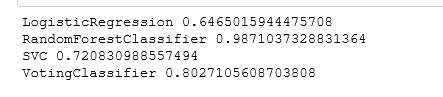
|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **ASE** |
| Simple Linear | 67.625 | 57.521 |
| Multiple Linear | 0.034 | 0.026 |
| Random Forest Regression | 0.0002 | 0.0002 |

For the binary classification analysis, I tested out several different algorithms which produced varying results. I trained the data on a Stoachastic Gradient Descent (SGD) algorithm and which did not perform very well,*(Cell 420)* producing cross validation scores of 0.66099311, 0.66038267, and 0.60489658 along with a precision score of .6462 and an f-score of .7073.

On the other hand, the random forest classifier produced very high accuracy rates, as shown in the chart below *(Cell 458)*



It would seem that the random forest classifier is the best performer. Would including it in an ensemble method result in even better performance? I tried it out by running a hard voting classifier using logistic regression, random forest, and SVC. *(Cell 439 ).*  As shown by the results below, the voting classifier performs worse than the Random Forest Classifier on its own.



**Running the Algorithms on the Test Set**

I ran the decision tree regressor on the test set and received an accuracy of .0002 *(Cell 932)*

I ran the Random Forest Classifier on the test set and received an accuracy of 0.995 *(Cell 447)*

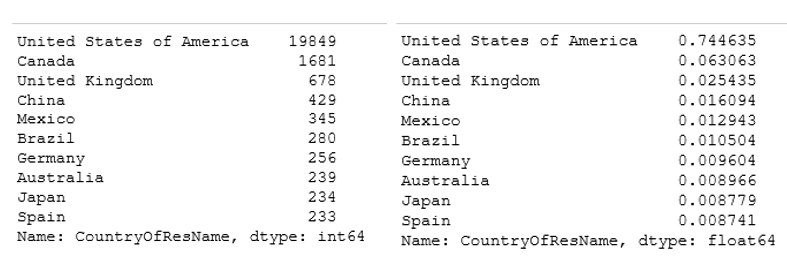
This indicates to me that the training set results are not overfitting the data.

**Going the Extra Mile: Additional Data Exploration on Which City/State/Country most runners come from and which country has the fastest runners**

To satisfy my curiosity about the geographic distribution of runners and relationship between nationality and finishing time, I performed some additional exploratory data analysis on the data set, bringing back features related to runners’ location. They included answers to the following questions:

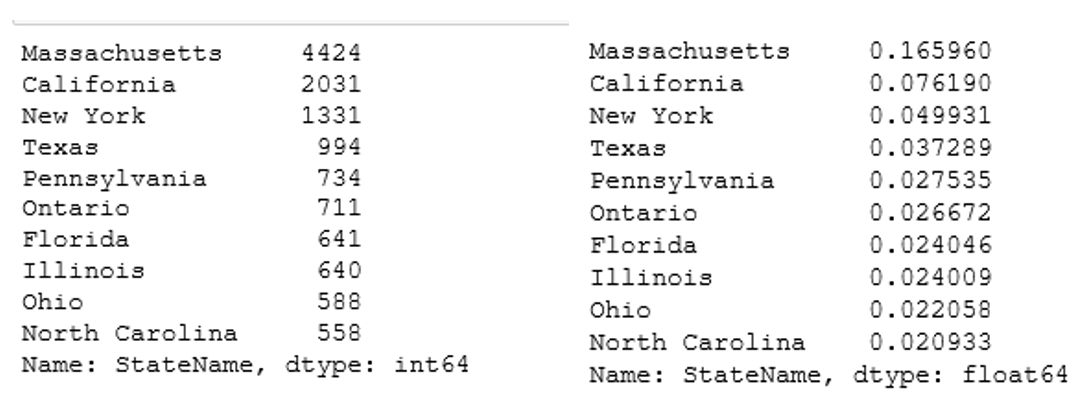
1. *Which countries have the most runners?*

A: Not surprisingly, the United States ranks first by far, followed by Canada and the other nations shown below *(Cell 511, 512):*



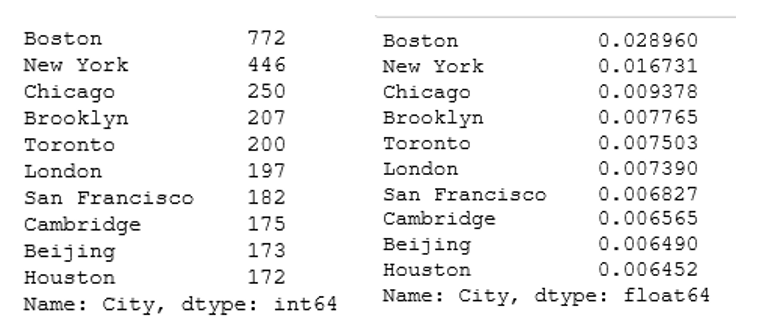
2. *Which states had the most runners?*

A: Massachusetts, the home state of the marathon, had the most runners, followed by US states with large overall populations and the Canadian Provence of Ontario *(Cell 933, 934):*



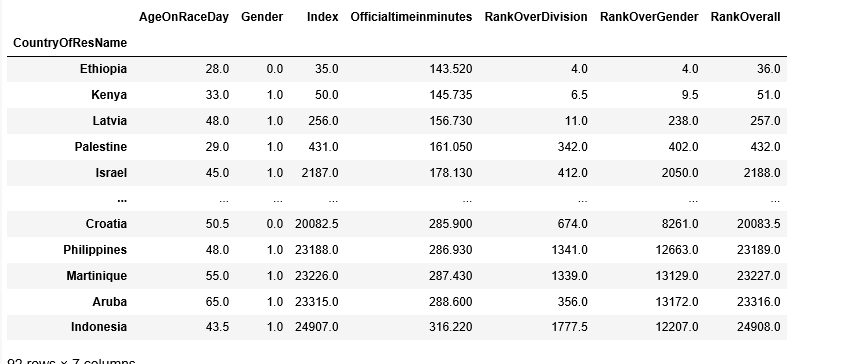
3. *Which city has the most runners?*

A: Boston has the most followed by large US and international cities (interestingly, Brooklyn is listed as a separate city whereas it is also a borough in New York) *(Cell 522, 42):*



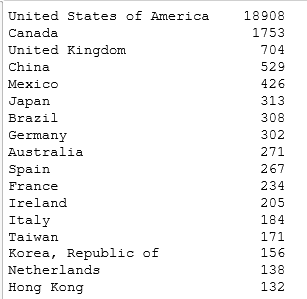
*4. Which country has the fastest runners?*

A: A pivot table organizing median finishing time by Country of Residence *(Cell 524)* shows the long distance running powerhouses of Ethiopia and Kenya with the fastest median times, followed by Latvia, Israel, and Palestine.

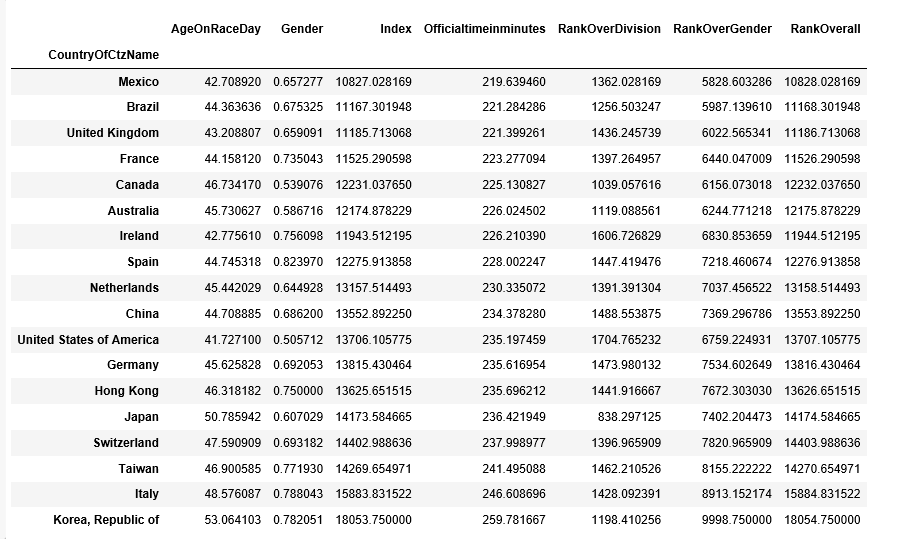


However, Latvia and Palestine only had one runner each and Israel only had seven runners *(Cell 547, 546, 549)*, so it appears that they ranked high by virtue of some very fast individuals.

It might make sense, instead, to compare finishing times amongst countries with 100 or more runners. The table below shows the countries with over 100 runners *(Cell 560)*:



When I ran a pivot table on the data set consisting of these countries, it looks like Mexico has the fastest runners, followed by Brazil, the United Kingdom, and France *(Cell 561):*



**Conclusions, Caveats, and Future Research**

At first glance, this paper merely confirms information that most people (even people who don’t subscribe to “Runner’s World”) already know: on average, runners get slower as they age and men generally run faster than women. What may be more interesting is how well the decision tree and random forest classifiers perform on identifying the gender of runners based on the data provided. While the scatter plot of men and women runners did not provide an obvious linearly severable data set, the decision tree and random forest binary classification algorithms performed with close to 100% accuracy.

That said, it would be a mistake to overgeneralize from the results of the 2019 Boston Marathon data to all marathons, or even future Boston Marathons. Finishing times can vary significantly depending on the weather, temperature, and other factors that change from year to year.

Future research could involve looking at a larger data set consisting of multiple marathon results as well as adding additional features such as bib number and some additional categorical data to the machine learning algorithm inputs.

1. The bibnumber could have been an interesting feature to have included in my machine learning analysis, had I had additional time. It would have required additional data cleanup, including transforming an object to an integer. [↑](#footnote-ref-1)
2. I tried hard to find a way to separate the charity runners from the runners who qualified based on prior marathon time but could not find information on the BAA website or elsewhere that would allow me to identify charity runners. Limiting my data exploration only to qualifying runners would reduce much of the variability in the exploratory analysis and machine learning results. [↑](#footnote-ref-2)
3. Had I had more time, I might have attempted to convert some of the categorical variables to numeric information using one-hot encoding in order to use them in machine learning algorithms. [↑](#footnote-ref-3)