Project 3 - London Marathon Predictions

William Sorg.74

2025-04-27

Question Formulation

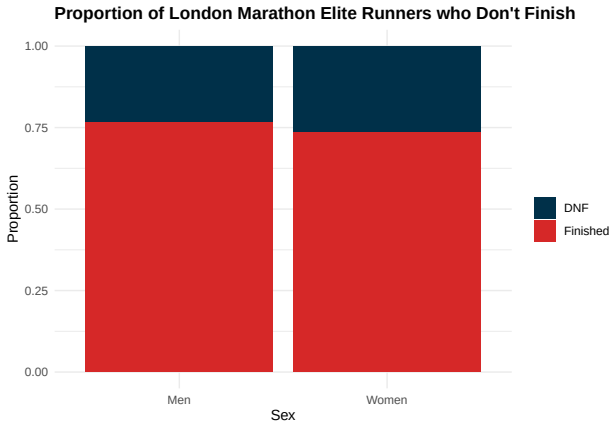
The London Marathon is one the most prestigious races in the world. As one of the Abbott World Marathon Majors, the London Marathon attracts over 50,000 runners to compete each year, along with approximately 750,000 spectators. Among the runners are some of the greatest elite distance runners in the world. Eliud Kipchoge, Mo Farah, Sifan Hassan, and the late Kelvin Kiptum have all raced the streets of London.
Despite the marathons's significance, it's difficult to find statistically driven predictions for the race. Nearly every other sport has some sort of win probability model, but there is a void for runnning. In a sport that can witness extreme unpredictability, I believe it would be interesting and informative to build a model that predicts the winner and top finishers of the London Marathon. Given that the 2025 London Marathon will begin at 4:00 a.m. EST on April 27th, this project will focus on predicting the results of the 2025 race. I will then briefly evaluate the accuracy of the predictions after the race concludes.

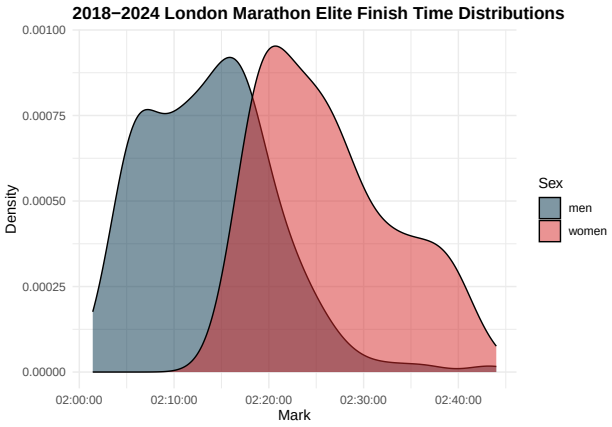
Data Selection

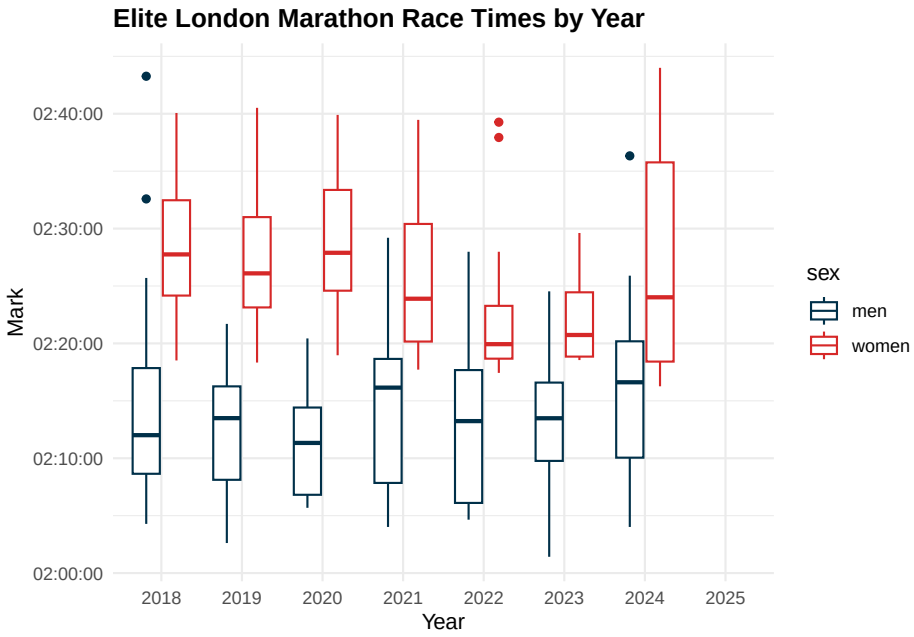
This project will use the 2018-2024 London Marathons to train and test different models with the goal of developing predictions for the 2025 London Marathon on April 27th. The London Marathon Elite runner results were scraped from each year's London Marathon results website. Then, for each athlete, every past race result was scraped off of the World Athletics website. From this data, several different statistics were calculated. These include number of races, number of finishes, PR, worst time, average time, time standard deviation, and most recent result for marathons, half marathons, and 10Ks. All code for scraping can be found on my GitHub. (Note, scraping the data was the biggest time requirement of this project.)

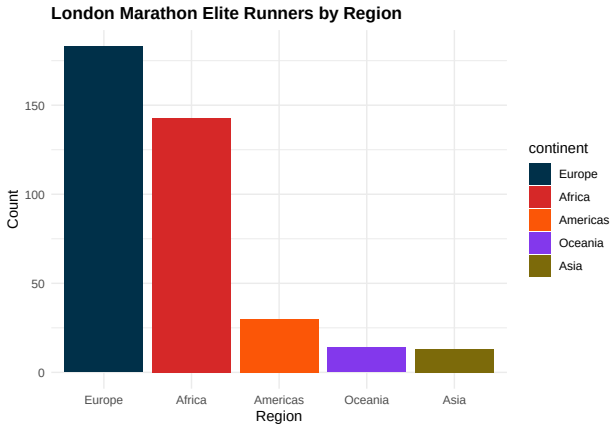
```
## # A tibble: 6 x 36
##   raceDate   Name             Nat   continent sex   age   Mark  Place dob
##   <date>     <chr>            <chr> <chr>     <chr> <drt> <Per> <int> <date>
## 1 2018-04-22 Guye Idemo ADOLA ETH   Africa    men   1004~ 9155S    17 1990-10-20
## 2 2018-04-22 Tracy BARLOW     GBR   Europe    women    N~ 9129S     9 NA
## 3 2018-04-22 Kenenisa BEKELE  ETH   Africa    men   1309~ 7733S     6 1982-06-13
## 4 2018-04-22 Tadelech BEKELE  ETH   Africa    women  987~ 8500S     3 1991-04-11
## 5 2018-04-22 Stephanie BRUCE  USA   Americas  women    N~ 9148S    10 NA
## 6 2018-04-22 Fernando CABADA  USA   Americas  men   1314~ 8259S    13 1982-04-22
## # i 27 more variables: lastRaceDate <date>, numMarathons <int>,
## #   finishesMarathons <int>, finishpctMarathons <dbl>, prMarathon <dbl>,
## #   worstMarathon <dbl>, avgMarathon <dbl>, sdMarathon <dbl>,
## #   recentMarathon <Period>, marathons <list>, numHalfMarathons <int>,
## #   finishesHalfMarathons <int>, finishpctHalfMarathons <dbl>,
## #   prHalfMarathon <dbl>, worstHalfMarathon <dbl>, avgHalfMarathon <dbl>,
## #   sdHalfMarathon <dbl>, recentHalfMarathon <Period>, ...
```

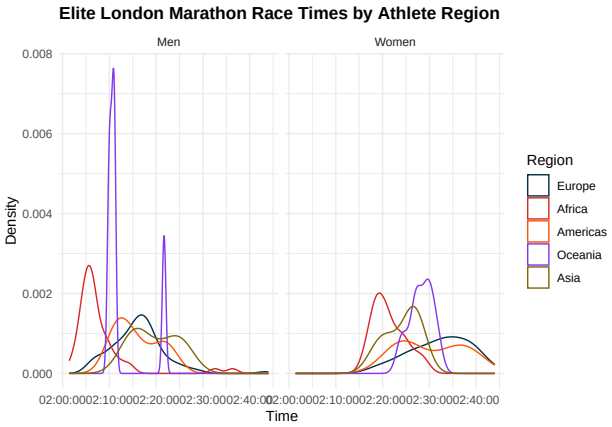
Exploratory Analysis

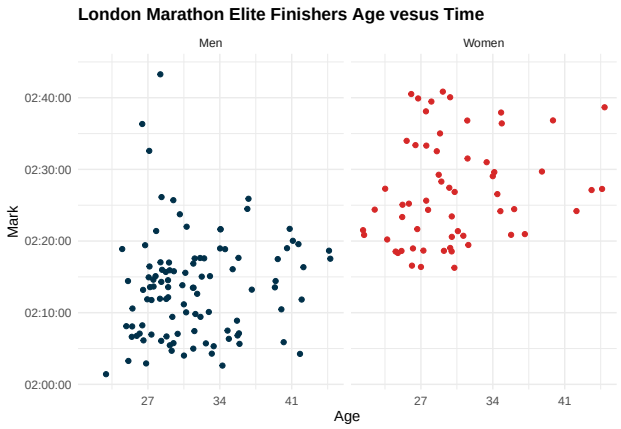


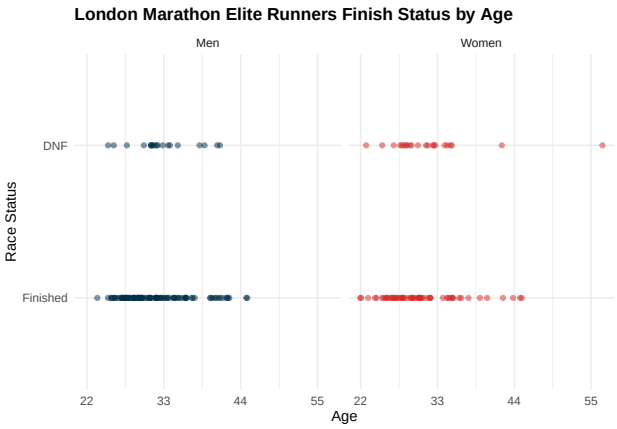


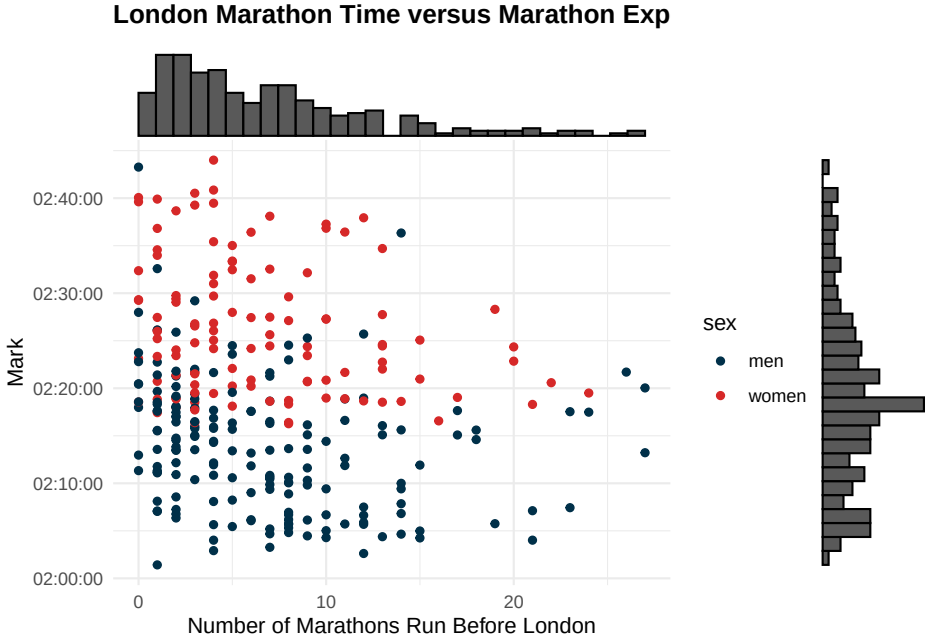


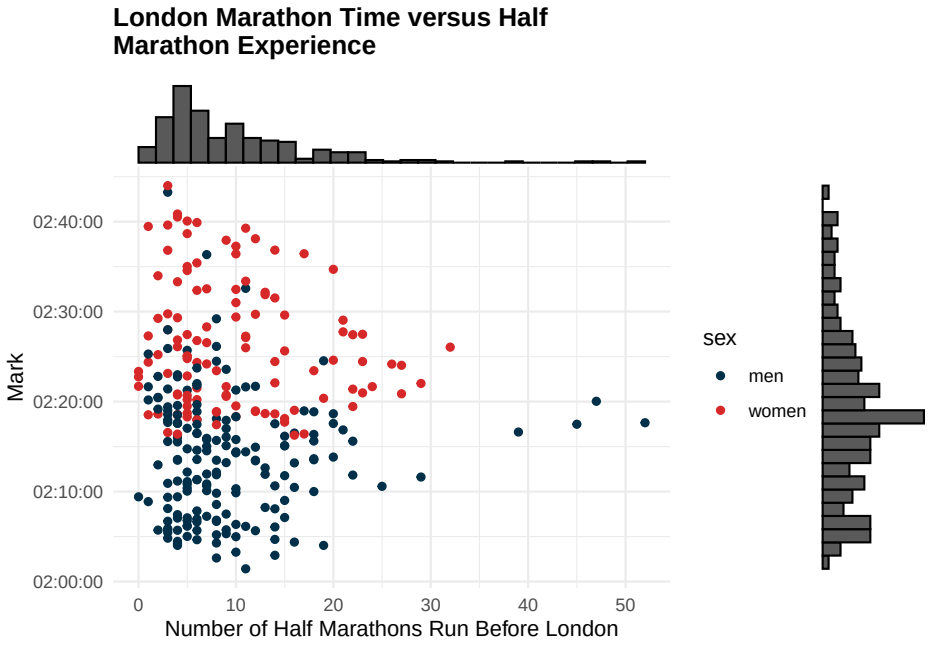


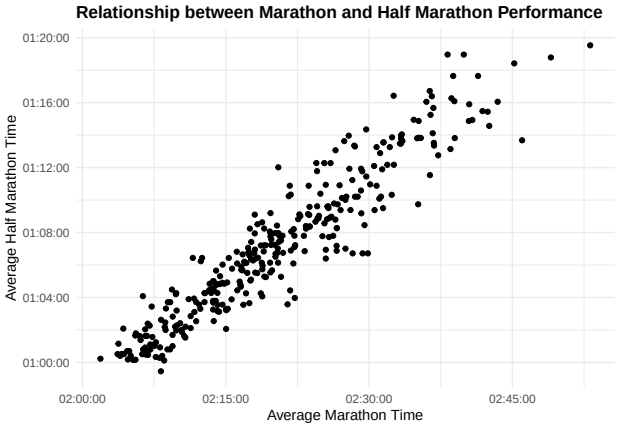


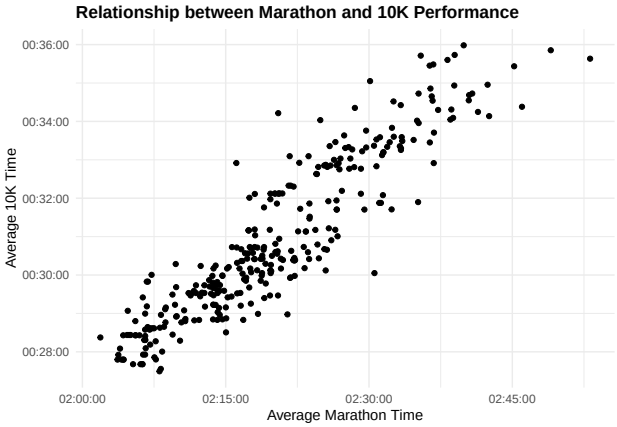

London Marathon Elite Finishers Age vesus Time
Men
Women
Mark
02:40:00
02:30:00
02:20:00
02:10:00
02:00:00
27
34
41
Age

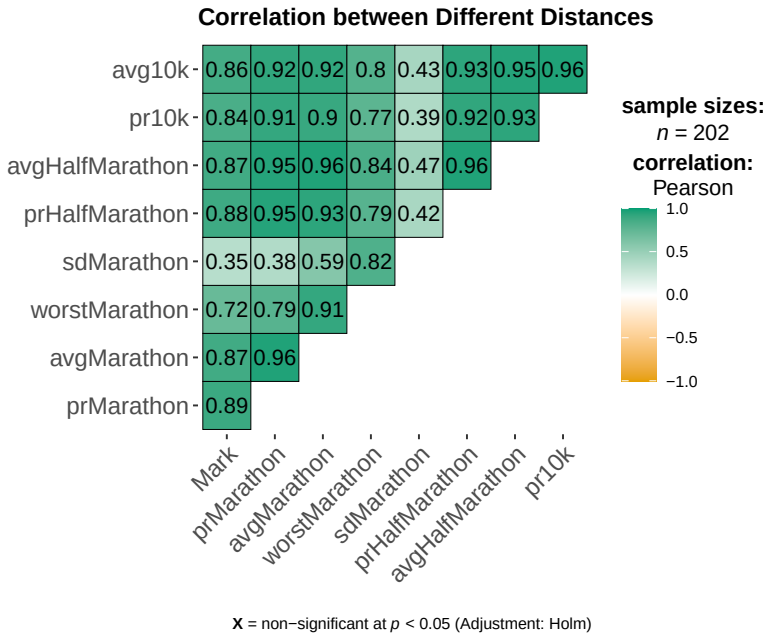

London Marathon Elite Runners Finish Status by Age
Men
Women
Race Status
DNF
Finished
22
33
44
55
Age


London Marathon Time versus Marathon Exp
Mark
02:40:00
02:30:00
02:20:00
02:10:00
02:00:00
0
10
20
Number of Marathons Run Before London
sex
men
women

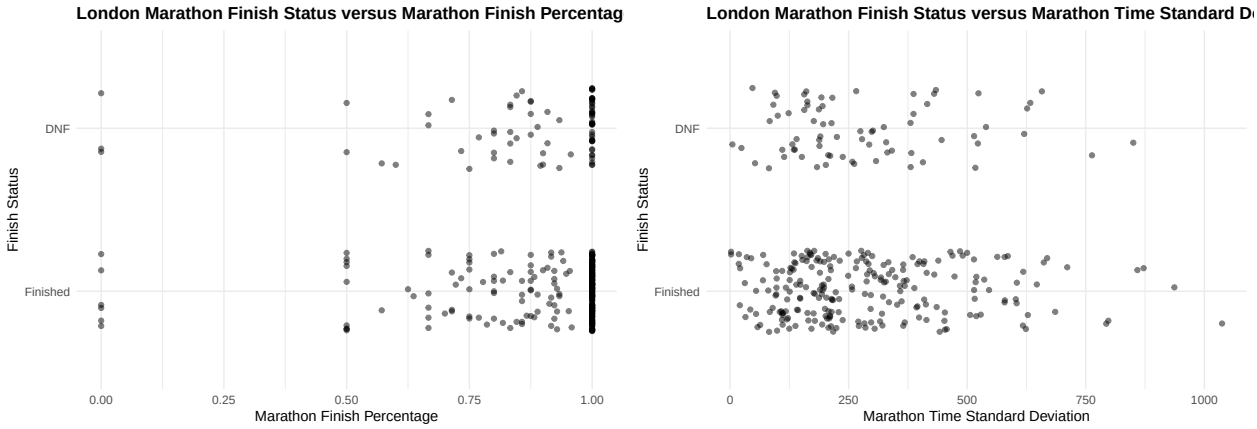

London Marathon Time versus Half Marathon Experience
Mark
02:40:00
02:30:00
02:20:00
02:10:00
02:00:00
0
10
20
30
40
50
Number of Half Marathons Run Before London
sex
men
women


Relationship between Marathon and Half Marathon Performance
Average Half Marathon Time
01:20:00
01:16:00
01:12:00
01:08:00
01:04:00
01:00:00
02:00:00
02:15:00
02:30:00
02:45:00
Average Marathon Time


Relationship between Marathon and 10K Performance
Average 10K Time
00:36:00
00:34:00
00:32:00
00:30:00
00:28:00
02:00:00
02:15:00
02:30:00
02:45:00
Average Marathon Time



X = non−significant at $p < 0.05$ (Adjustment: Holm)



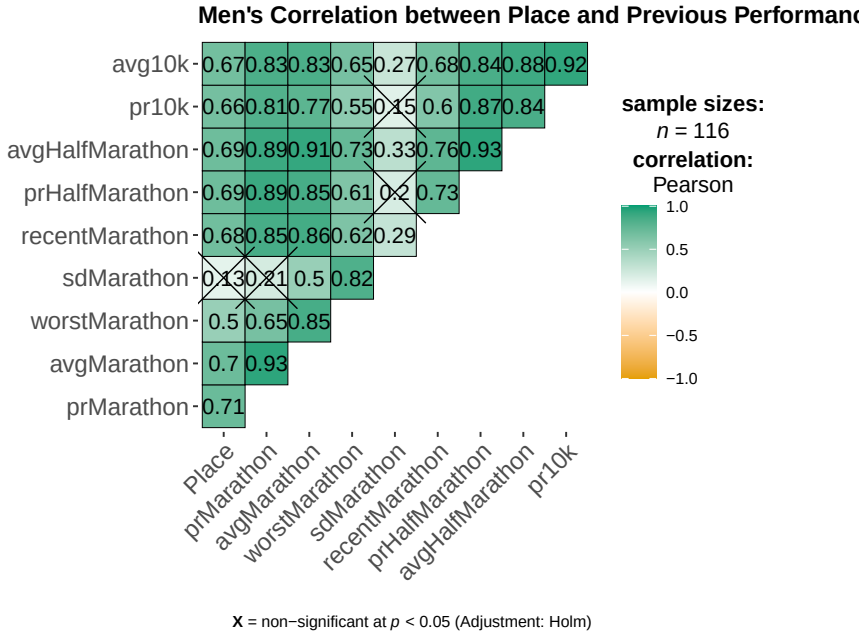

```
t.test(LMAnalysisData$finishpctMarathons[is.na(LMAnalysisData$Mark)], LMAnalysisData$finishpctMarathons
```

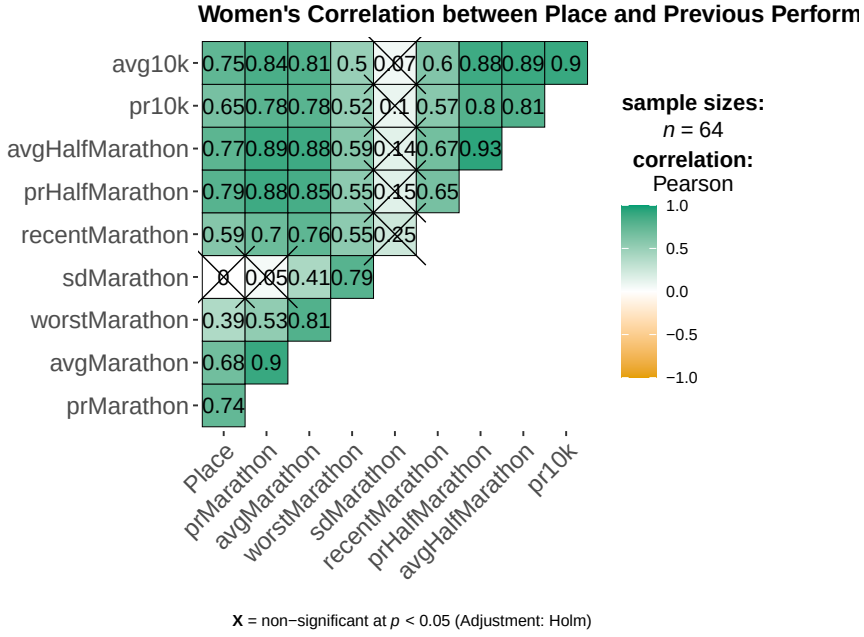
```
##
##  Welch Two Sample t-test
##
## data:  LMAnalysisData$finishpctMarathons[is.na(LMAnalysisData$Mark)] and LMAnalysisData$finishpctMara
## t = -1.2083, df = 122.68, p-value = 0.2293
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08205179  0.01985097
## sample estimates:
## mean of x mean of y
## 0.8808762 0.9119766
```

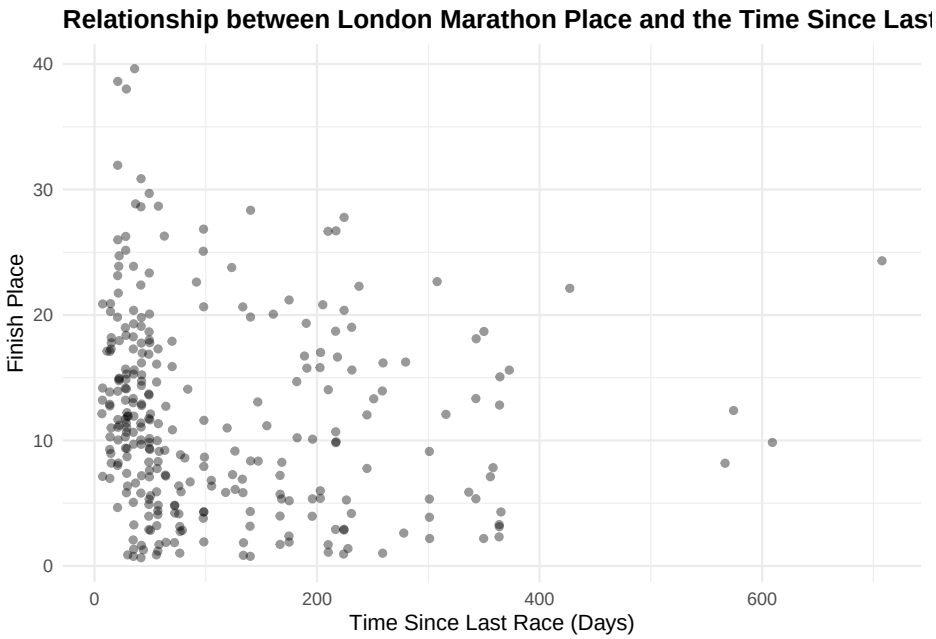
```
t.test(LMAnalysisData$sdMarathon[is.na(LMAnalysisData$Mark)], LMAnalysisData$sdMarathon[!is.na(LMAnalysi
```

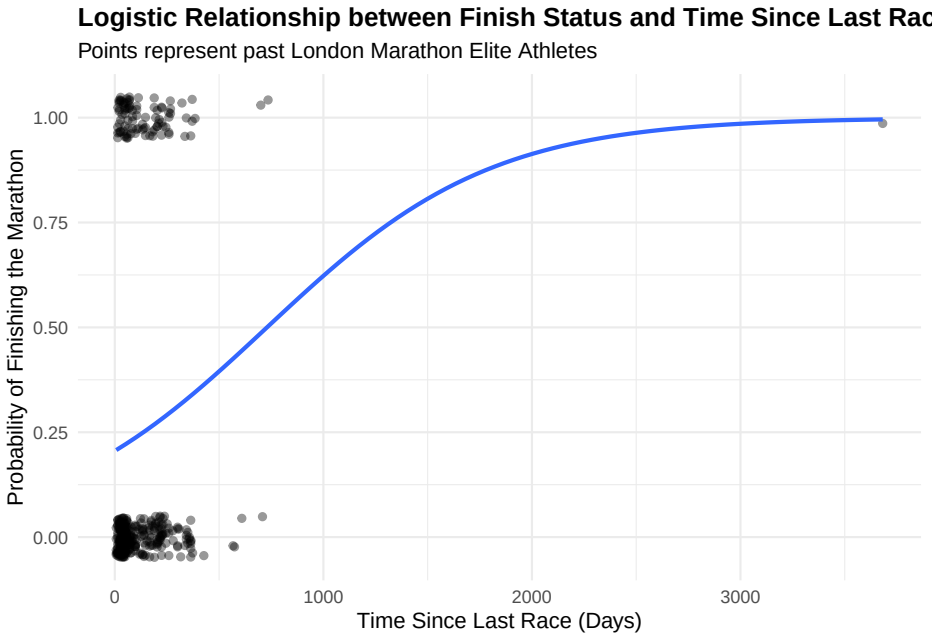
```
##
```

```
##  Welch Two Sample t-test
##
## data:  LMAnalysisData$sdMarathon[is.na(LMAnalysisData$Mark)] and LMAnalysisData$sdMarathon[!is.na(LM
## t = -0.31434, df = 126.2, p-value = 0.7538
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -56.34319  40.89727
## sample estimates:
## mean of x mean of y
##  279.6781  287.4011
```









Analysis

Recommendations