Project1 Codebook

Yingxi Kong

2024-10-01

Abstract

Introduction

Marathon running, a long-distance race covering 42.195 kilometers (26.2 miles), has been known for its physical and mental challenges. Over the years, it has attracted an increasing number of excellent runners from around the world, leading to interests in how physiological and environmental conditions influence runner's performance. Understanding these factors would help runners to better understand their performance. Also, it would provide more information to the runners to optimize their training plans and race-day strategies, leading to improvement of their performance.

This report is a collaboration with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College, which explores how environmental conditions, age, and sex would influence runner's performance in this long-distance race. Their prior research found that warmer temperature leads to decline in performance in marathon races, and this decline in endurance performance varies significantly between females and males. Moreover, older adults face more thermoregulatory challenges during exercise, which further exacerbate performance declines under warmer temperature. This exploratory analysis study aims to build on previous findings, providing deeper insight by investigating the intersection of age, sex, and environmental conditions on runners' marathon performance.

Data and Preprocessing

The primary dataset in this project consists of 11,564 observations with 14 variables, including information on participants' characteristics and environmental condition characteristics collected from five major marathon races in the U.S.: the Boston Marathon, Chicago Marathon, New York City Marathon, Twin Cities Marathon (Minneapolis, MN), and Grandma's Marathon (Duluth, MN). This data spans a period of 15-20 years. Our response variable is `CR_pct`, representing the percent off the current course record. Key covariates include race, sex, temperature (`Tdc`, `Twc`, `Tgc`, etc.), relative humidity (`rh`), solar radiation (`SRWm2`), dew point (`DP`), wind speed (`Wind`), and Wet Bulb Globe Temperature (`WBGT`).

There are 491 observations with missing values and missingness appears mainly among environmental conditions covariates from races held in 2011 and 2012. These missing observations were neither excluded from the analysis nor imputed to retain as much data as possible for exploration.

During the preprocessing steps, we managed the column names for easier understanding, modified columns with coding issues, and ensured the correct data types for various columns. In addition, we merged additional datasets from multiple sources with our primary data, including the `course_record` dataset, which contains the best course record times for each race by gender, and the `aqi_values` dataset, which includes information of air quality for each race day through years.

Moreover, we mutate a new column `age_group` for further aging analysis. Based on the age distribution in the dataset, runners are classified into five aging groups: Younger, Lower-Mid Age, Mid Age, Upper-Mid Age, and Highest Age with cut-off points presented in Table 1. These cut-off points are decided by considering

both the breaks in the data and common practices in marathon age classifications. Instead of using the traditional 10-year groups, we set 15-year groups to better reflect the age distribution of participants in our dataset.

Table 1: Age Group Classifications for Marathon Participants

Group	Age Range
Younger	Below 25
Lower-Mid Age	25-39
Mid Age	40-54
Upper-Mid Age	55-69
Highest Age	70+

Exploratory Data Analysis

The summary statistics in Table 2 provides a comprehensive understanding of environmental conditions experienced during the five major marathons. The Boston Marathon was the only rave with a completed record of weather conditions. Observing the pattern of weather conditions, the Grandma's Marathon consistently exhibited the warmest temperature where they had the highest average dry bulb temperature ($18.9°C$), wet bulb temperature ($14.9°C$), and black global temperature($32°C$), reflecting a higher risk of heat illness. Moreover, the Grandma's Marathon had 47% of time falling within the WGBT range of $18 - 23\ °C$. In contrast, the Boston and New York City Marathons had the coolest temperature with more races falling within the WGBT range of $< 10\ °C$. Similarly, these two Marathons presented lowest average dry bulb, wet bulb, and black global temperatures. The Chicago and Twin Cities Marathons had relatively moderate temperature with more races falling within the WGBT range of $10 - 18\ °C$. Wind speeds were highest in Boston and New York City and remain consistent around 9 km/h for other races. Boston Marathon exhibited highest air quality index (`aqi`) and solar radiation level (`SRWm2`) while New York City Marathon exhibited the lowest for both. Moreover, Chicago Marathon had the highest level of relative humidity and New York City Marathon had the lowest. Based on this summary statistics, we observed the significant variation in environmental conditions across different location which might have a substantial impact on the runners' performance.

We first investigate the relationship between age and performance by gender. Table 3 presents the summary of participants' performance based on their age group and gender. The younger and highest age groups have fewer observations due to the limited number of participants at these extreme ages. The lower-mid age group participants had the best performance, with both genders exhibiting the lowest %CR values across all aging groups. People with younger, Mid, and upper-mid age have relatively worse performance compared to lower-mid age people. Older people (Highest age) show the largest deviation from the course record, with an average %CR at around 130-140. In addition, female runners generally have worse performance (higher %CR value) compared to males runners within the same age range.

Additionally, in Figure 1, both genders' plots exhibit a U-shaped curve which further explains the relationship between age and performance. The performance continuously improves as age increases (%CR decreases as age increases), achieve the optimal performance (lowest %CR), and then gradually declines as age increases (%CR increases as age increases). Both gender reach their optimal performance during their Lower-Mid age where female participants' average %CR achieves the lowest value at age of 28 and male participants achieves the lowest average %CR at age 30. After reaching their optimal performance, both genders exhibits decline as age increases. Male participants show more extreme decline compared to female participants especially during their upper-mid and highest age, end with a maximum average percent off course record (CR%) exceeding 300.

Table 2: Summary Statistics by Race

Characteristic	Race				
	Boston Marathon, N = 18[1]	**Chicago Marathon**, N = 21[1]	**Grandma's Marathon**, N = 17[1]	**New York City Marathon**, N = 23[1]	**Twin Cities Marathon**, N = 17[1]
flag					
Missing	0 (0%)	1 (4.8%)	1 (5.9%)	1 (4.3%)	1 (5.9%)
WBGT < 10C	9 (50%)	6 (29%)	0 (0%)	11 (48%)	5 (29%)
WBGT > 18-23C	1 (5.6%)	1 (4.8%)	8 (47%)	4 (17%)	3 (18%)
WBGT > 23-28C	1 (5.6%)	1 (4.8%)	2 (12%)	0 (0%)	1 (5.9%)
WBGT 10-18C	7 (39%)	12 (57%)	6 (35%)	7 (30%)	7 (41%)
Dry bulb temperature	11.6 (6.0)	12.4 (6.2)	18.9 (3.4)	11.7 (4.8)	13.2 (5.7)
Missing	0	1	1	1	1
Wet bulb temperature	7.6 (3.9)	8.6 (5.9)	14.9 (2.5)	7.6 (5.1)	9.9 (5.6)
Missing	0	1	1	1	1
Percent relative humidity	35 (35)	61 (11)	49 (36)	27 (31)	41 (35)
Missing	0	1	1	1	1
Black globe temperature	24 (9)	25 (6)	32 (8)	21 (6)	25 (7)
Missing	0	1	1	1	1
Solar radiation in Watts	654 (191)	460 (96)	679 (195)	401 (134)	437 (143)
Missing	0	1	1	1	1
Dew Point	3 (5)	5 (7)	12 (3)	3 (7)	6 (8)
Missing	0	1	1	1	1
Wind	12.0 (4.6)	8.2 (3.3)	9.2 (2.9)	11.2 (4.7)	8.8 (3.3)
Missing	0	1	1	1	1
WBGT	11.3 (4.6)	12.1 (5.9)	18.6 (3.3)	10.7 (5.0)	13.3 (5.6)
Missing	0	1	1	1	1
Air Quality Index	42 (15)	40 (13)	37 (15)	33 (14)	35 (15)

[1] Mean (SD) for continuous; n (%) for categorical

From these results, we found that age does affect the marathon performance for both genders. Runners' performance gradually enhances from their younger age to lower-mid age, reaching their peak at age around 30, after which performance gradually declines. There are also gender differences that females usually perform worse compared to males within the same age group. In addition, male runners with higher age have more pronounced decline in performance compared to female runners.

Environmental factors such as humidity, WBGT, and wind speed are also suspected to be key influences on marathon performance. To investigate the effects of these environmental condition characteristics on marathon performance, we first examine the correlations between key environmental variables and performance metrics by the correlation plot in Figure 2. Strongly correlated variables, either positively or negatively, are in blue, and weakly correlated variables are presented in white. Observing Figure 2, WBGT, dry bulb, wet bulb, and

black global temperature show strong positive correlation with each other. This is reasonable since WBGT is calculated with those temperature variables using the following formula:

$$WBGT = (0.7 * Twc) + (0.2 * Tgc) + (0.1 * Tdc)$$

All other environmental characteristics have moderate or week relationship among each other. For example, the relative humidity (`rh`) is weakly and negatively correlated with `Tdc` and `Tgc` with correlation value of -0.01.

Figure 1: Overall Performance vs. Age by Gender

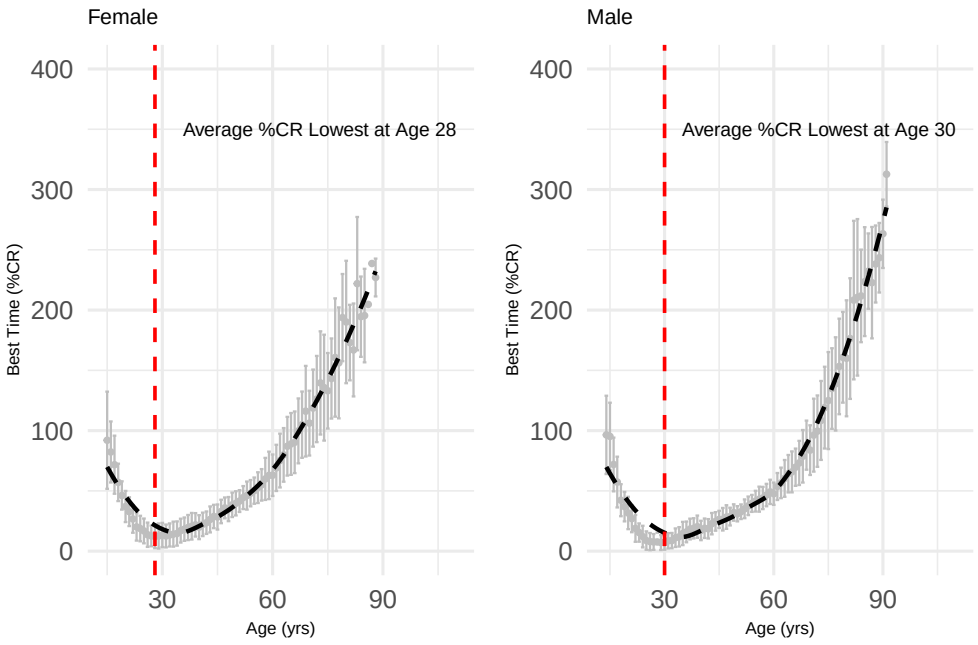
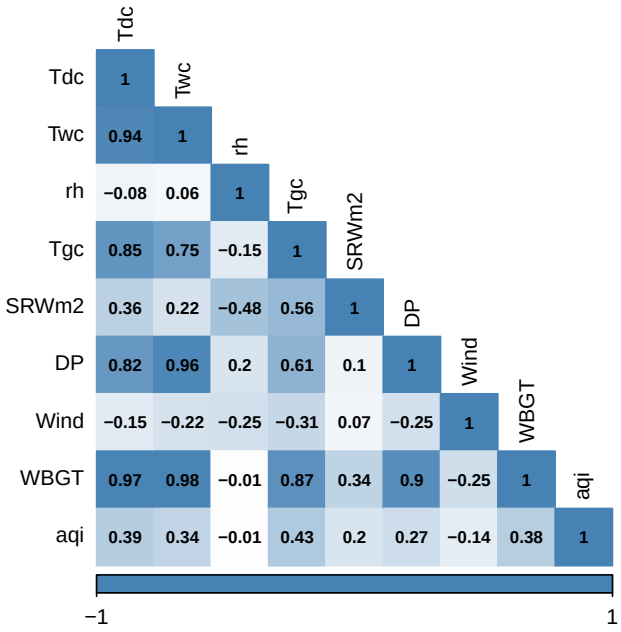


Table 3: Summary of Marathon Performance by Age Group and Sex

Age Group	Sex	N	Min Performance	Mean Performance	Median Performance	Max Performance
Younger	Female	788	-0.385	40.840	38.051	211.095
Younger	Male	834	-1.074	36.119	31.196	159.535
Lower-Mid Age	Female	1440	-1.816	15.547	14.218	50.301
Lower-Mid Age	Male	1440	-2.251	11.645	9.322	50.179
Mid Age	Female	1440	-1.419	34.064	33.975	76.215
Mid Age	Male	1440	1.243	27.992	28.445	89.271
Upper-Mid Age	Female	1346	21.154	75.345	69.512	273.824
Upper-Mid Age	Male	1435	11.663	58.746	55.054	153.190
Highest Age	Female	438	36.187	138.609	129.911	336.347
Highest Age	Male	963	48.375	132.480	120.909	419.958

As mentioned previously, `WBGT` is calculated using all temperature variables and it is highly correlated with those variables. The strong correlation between `WBGT` and these variables suggests that `WBGT` alone can be a sufficient indicator of how temperature characteristics influence marathon outcomes. Figure 3 presents an illustration of how `WBGT` effects marathon performance across different age groups and genders, using `geom_smooth` to show the performance trend. Higher `WBGT` values indicate higher risks of heat stress illness.

Figure 2: Correlation Plot among Environmental Condition Characteristics



Runners in higher, upper-mid, and Younger age groups are more susceptible to the effects of increasing `WBGT` that their performance fluctuates more as `WBGT` increases. Additionally, as `WBGT` increases, runners' performance generally worsens across all groups (%CR increase). Runners gradually perform better as `WBGT` value increase from $0°C$ to $10°C$, but their performance then declines when `WBGT` exceeds $10°C$. For older runners in the highest age group, this decline is more significant, leading to largest deviation from the course record. Additionally, male runners consistently perform better compared to female runners within same age group as we mentioned earlier. Male runners in the younger, upper-mid, and highest age group show steeper decline in performance when `WBGT` is larger than $10°C$ compared to female younger runners. Although male runners perform better generally, they are more sensitive to higher `WBGT` levels in specific age groups.

Figure 3: Runners' Performance vs. WBGT

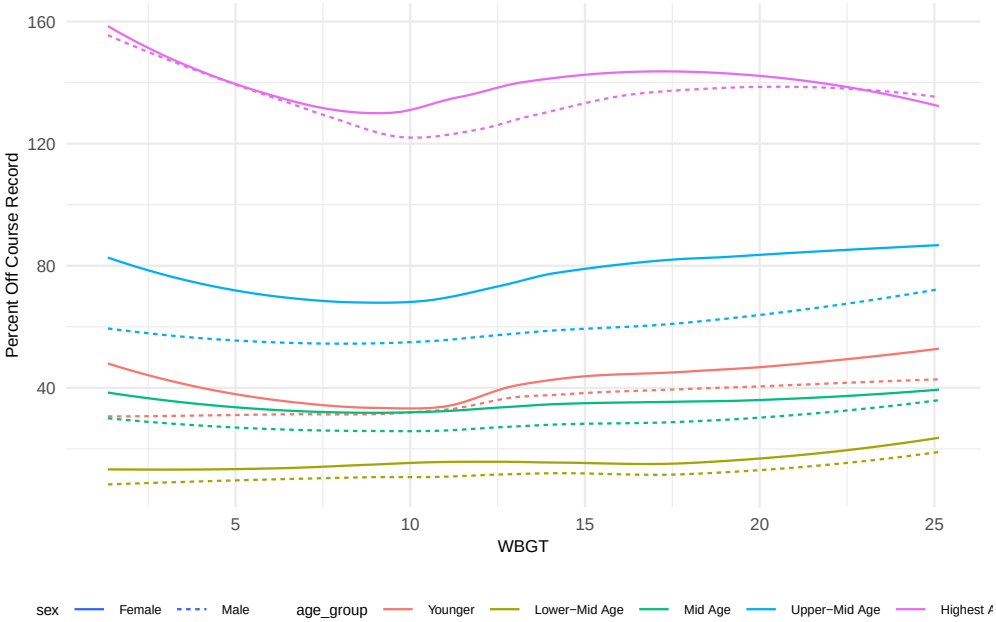


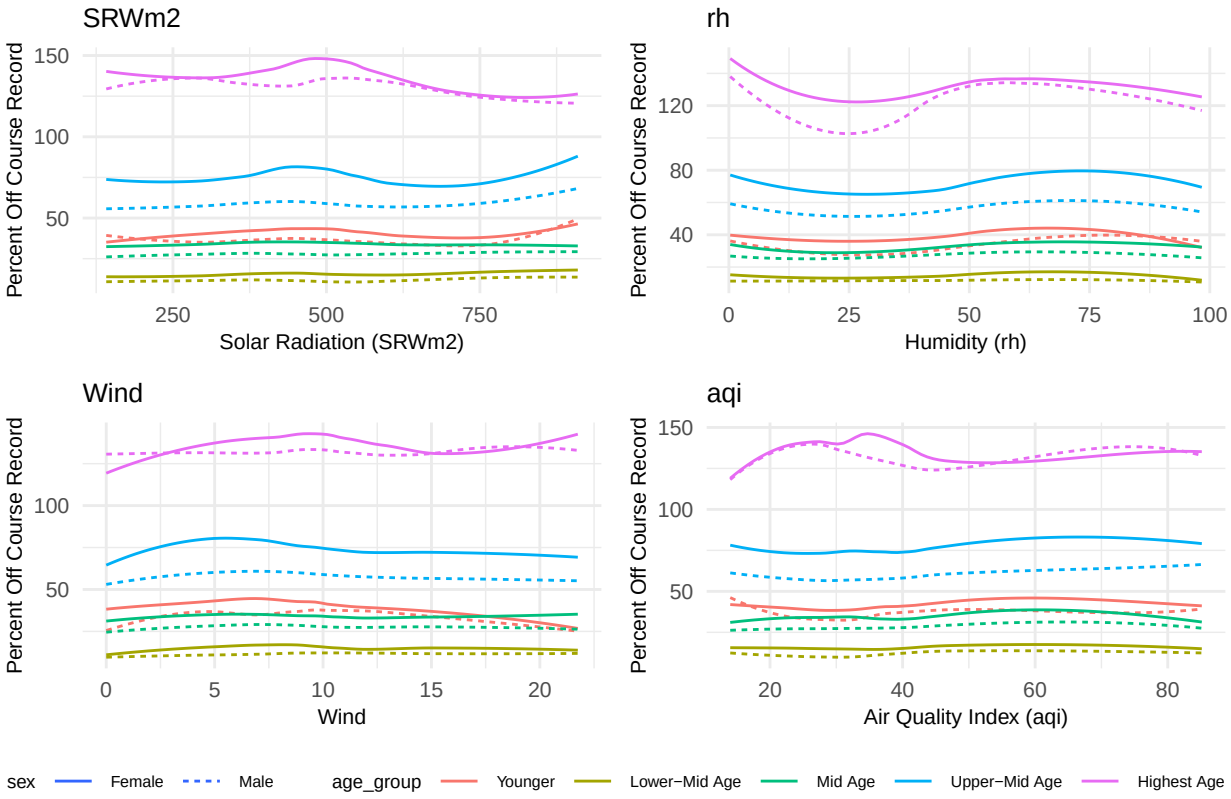
Figure 4 presents the relationship between other environmental condition characteristics and runners' performance across different genders and age groups.

Starting from Solar radiation (`SRWm2`), runners in the lower-mid and mid age groups seem to maintain a more stable performance as `SRWm2` increase while younger, upper-mid, and highest age runners are more sensitive to the level of solar radiation, with their performance fluctuating more as solar radiation level increases. Moreover, female runners across groups show more stable performance compared to male runners that female runners are less likely to be influenced by solar radiation level.

People in the younger, upper-mid, and highest age groups still show more fluctuated trend in relative Humidity (`rh`) as well. At lower humidity levels from 0 to 30%, runners across groups have improvement as humidity level increase. This indicates that moderate humid conditions would not negatively influence runner's performance but help runners by preventing dehydration and excessive sweating. However, as humidity rises excess this range, runners start to have worse performance since extremely high humidity level would makes runners maintain higher body temperature by preventing runners' body to sweating and cooling itself. Moreover, male runners in the highest age group are much more sensitive to the humidity change compared to female runners.

Wind speed (`Wind`) does not exhibits significant trend on runners' performance across various age groups and genders. For most age groups, %CR shows only minor changes as wind speed increases, suggesting that wind speeds do not have significant impact on runners' performance that runners are able to adjust under various wind conditions and maintain their great performance. Male runners in the highest age group show relatively greater fluctuation since older runners would be more sensitive to environmental condition change. Moreover, again, female runners present more stability in performance as wind speed varies as well.

Figure 4. Runners' Performance vs. Other Environmental Condition Factors



Finally, runners show consistent performance as air quality index increases, except for the highest age group which experienced relatively more pronounced variation for both genders. In addition, there are not obvious differences on variations between genders.

Thus, environmental condition characteristics like WBGT, solar radiation, and relative humidity do affect runner's performance. Older runners (upper-mid and highest age group) and younger runner (younger age group) are generally more sensitive to the environmental change. Females usually maintain higher stability through those environmental variations. Although wind speed and air quality do not present a pronounced effect on runners' performance, older runners still exhibits more sensitivity to these factors, while female runners continue to maintain more consistent performance compared to male runners.

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: CR_pct ~ rh + SRWm2 + Wind + WBGT + aqi + (1 | age)
##    Data: data
##
## REML criterion at convergence: 96888.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.2548 -0.4716 -0.0731  0.4160  9.2257
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  age      (Intercept) 5416     73.59
##  Residual              350     18.71
## Number of obs: 11073, groups:  age, 78
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  7.553e+01  8.400e+00  7.820e+01   8.992  1.1e-13 ***
## rh          -1.294e-02  6.615e-03  1.099e+04  -1.956   0.0505 .
## SRWm2       -3.558e-03  1.181e-03  1.099e+04  -3.012   0.0026 **
## Wind        -3.443e-02  4.689e-02  1.099e+04  -0.734   0.4628
## WBGT         4.659e-01  3.754e-02  1.099e+04  12.410  < 2e-16 ***
## aqi          1.002e-02  1.353e-02  1.099e+04   0.741   0.4587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) rh     SRWm2  Wind   WBGT
## rh    -0.070
## SRWm2 -0.059  0.491
## Wind  -0.076  0.199 -0.049
## WBGT  -0.025 -0.105 -0.328  0.226
## aqi   -0.036 -0.033 -0.096  0.059 -0.310
```