

Impact of Environmental Conditions on Maraton Runners' Performance Based on Gender and Age

Diahmin Hawkins

October 6, 2024

Introduction

In recent years, marathon participation and performance have seen a marked increase, prompting a deeper exploration into the factors influencing outcomes in these endurance events. In collaboration with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College, this study aims to assess the impact of environmental condition like temperature, humidity, solar radiation, and wind speed on marathon performance in both male and female marathon runners.

This study will focus on three aims. The first aim is to examine the effects of increasing age on marathon performance in men and women. Our second aim is to explore the impacts of environmental conditions on marathon performance, and whether the impact differs across age and gender. The last aim is to identify (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance. I hypothesize that increasing environmental temperatures and unfavorable weather conditions will have a negative impact on marathon performance.

Methods

Preprocessing

The initial exploratory analysis was conducted to identify patterns and relationships among key variables. During preprocessing, it was observed that **Race** (0=Boston, 1=Chicago, 2=NYC, 3=TC, 4=D) corresponds to the **Race** variable in the *course_record* dataset, while **Sex** (0 = F, 1 = M) corresponds to the **Gender** variable. To ensure consistency, the **Sex** and **Race** variables were renamed to **Gender** and **Race**.

Further review revealed discrepancies in how some marathon races were coded. For instance, while one dataset used numeric codes (e.g., 0 for Boston), another dataset used letter codes (e.g., B for Boston). To standardize these variables, the race codes were recoded from 0 to B, 1 to C, 3 to TC, and 4 to D. Additionally, the binary outcomes for **Gender** were modified to align across datasets, with 0 recoded as M for male and 1 as F for female.

Variable names were then systematically renamed for clarity and consistency with the codebook. After completing these preprocessing steps, the data were merged by **Race**, **Gender**, and **Year**, ensuring that variables from the left dataset were retained, while integrating other variables that were not common across datasets. To ensure precise measurement of Run Times, the initial step involved converting the course records to seconds. A new variable was established to calculate the actual runtime and marathon duration for each participant in seconds, defined by the formula: $\text{Runtimes} = \text{Race_Seconds} * (1 + (\text{Percent CR} / 100))$.

To begin with, we will observe the data and notice any commonalities within the variables. We begin our data preprocessing by noticing that **Race** (0=Boston, 1=Chicago, 2=NYC, 3=TC, 4=D) corresponds to the **Race** variables in the *course_record* data set.

Missingness

The raw data used for this analysis consisted of 11,564 rows and 14 columns. To begin this analysis, I got the sum of missing values from the dataset by columns. From this analysis, it was observed that the variables Flag, Dry Bulb Temp C, Wet Bulb Temp C, Percent Relative Humidity, Black Globe Temp C, Dew Point in C, Wind Speed, and Wet Bulb Globe Temp contained missing data. The missing data were examined using the `naniar` package in R to determine the percentage missing and available in the data. Following this procedure, each of these columns was found to have 491 missing values, totaling 4,419 missing values overall, with a 4.25 missingness percentage. To further quantify the extent of missingness, the `naniar` package in R was employed to calculate the percentage of missing and available data. The missing data represent only 2.2% of the dataset, while 97.2% of the data remains well-represented. Therefore, these missing data properties were removed from further analysis.

Missing Data Summary for Environmental Variables

| Variables | Missing Values | Percentage Missing (%) |
|---------------------------|----------------|------------------------|
| Flag | 491 | 4.25 |
| Dry bulb Temp C | 491 | 4.25 |
| Wet bulb Temp C | 491 | 4.25 |
| Percent Relative Humidity | 491 | 4.25 |
| Black Globe Temp C | 491 | 4.25 |
| Solar Radiation | 491 | 4.25 |
| Dew Point in C | 491 | 4.25 |
| Wind Speed | 491 | 4.25 |
| Wet Bulb Globe Temp | 491 | 4.25 |

The preprocessing of the marathon dataset and air quality involved several steps to ensure data compatibility. Therefore, race names were standardized and column names adjusted in the `marathon_dates` dataset to align with those in the `course_record_project1` data frame. Dates were formatted to (“YYYY-MM-DD”) to facilitate time-based operations and comparisons and the cleaned `marathon_dates` data was merged with the `course_record_project1` using a left join based on `Race` and `Year`.

The `aqi_values` data was updated to maintain consistency with the marathon data, standardizing race names and converting dates to a uniform format. Unnecessary columns were removed, and the average ozone concentration, measured over an 8-hour period, was calculated and summarized by `Race`, `Year`, and `Date`. Finally, this processed air quality data was merged with the `course_record_project1` data frame, combining marathon and air quality data to further analyze the impact of air quality on marathon performances.

Results

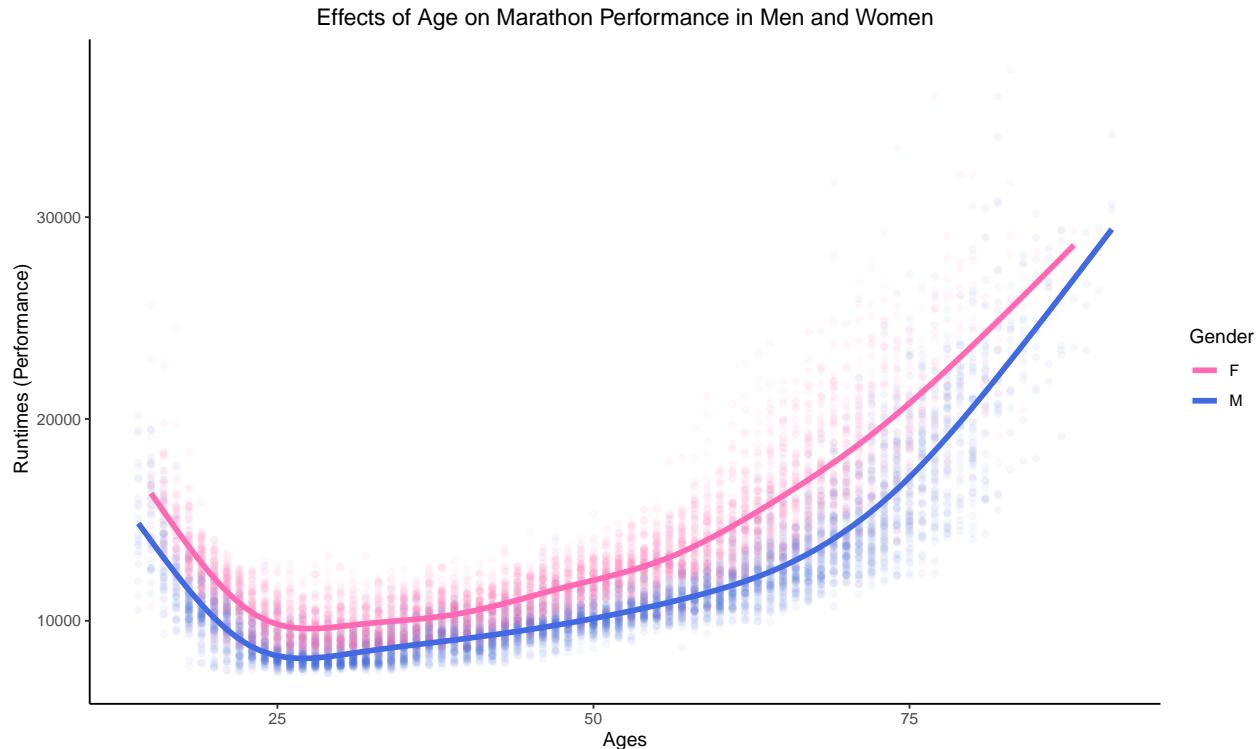
Aim 1: Examine effects of increasing age on marathon performance in men and women.

In the **Effects of Age on Marathon Performance in Men and Women** visualization, we observe the marathon performance (measured in runtimes) of men and women across a wide age range, from 14 to 91 years. Based on the data, there is a clear decrease in runtimes up to approximately 25 years of age, indicating improved performance in both men and women during their younger years, with peak performance typically occurring in the mid-twenties. This pattern signifies that younger participants, particularly those in their mid-twenties or younger, perform better in marathons compared to older participants.

After reaching this peak, there is a noticeable and steady increase in runtimes, suggesting a decline in performance as age advances. When comparing genders, men consistently demonstrate faster runtimes across

all ages. While both men and women exhibit similar trends in declining performance with age, the slope of decline is steeper for women, particularly around the ages of 60-65. This indicates that the negative effects of aging on marathon performance are more pronounced in women during later life stages.

Overall, the performance gap between genders widens with increasing age, with men generally maintaining faster runtimes compared to women, especially in older age groups. This widening gap highlights the greater impact of aging on female marathon runners in terms of performance decline.



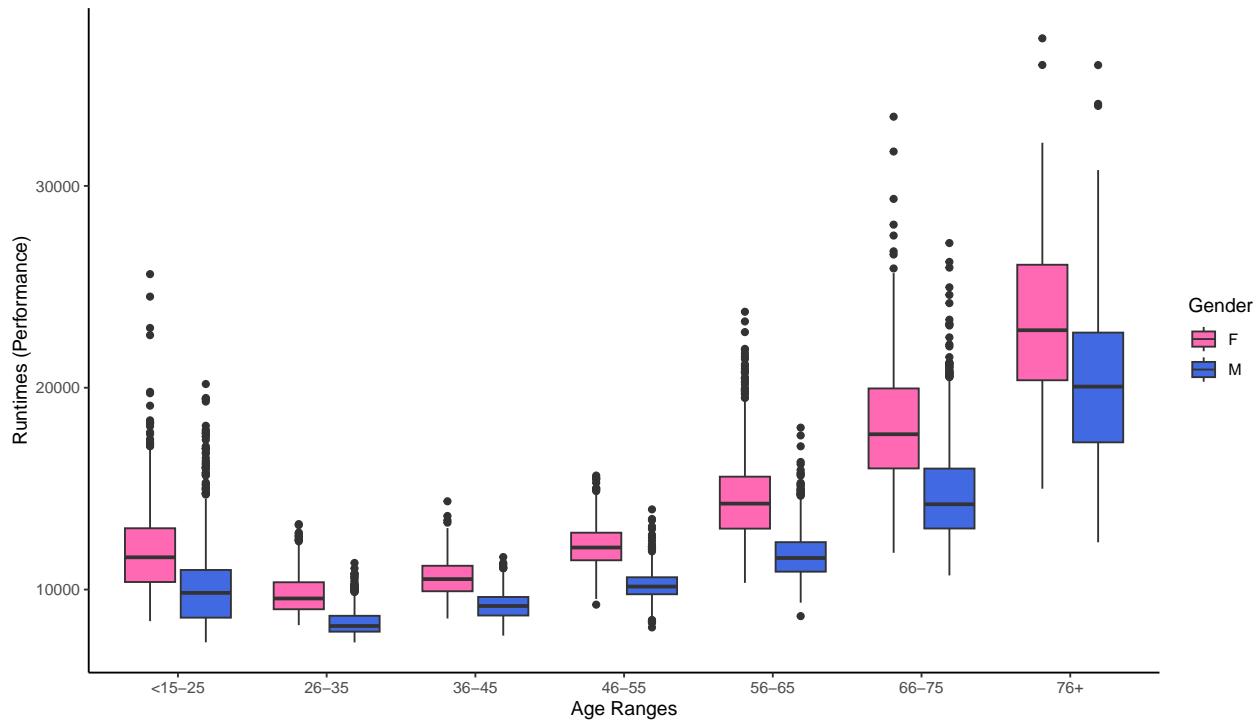
In the **Marathon Performance Summary Table by Age Ranges** and the **Effects of Age on Marathon Performance in Men and Women**, we noticed some summary statistics of the boxplot measurements of the marathon runners runtimes based upon their age ranges. The marathon runners were broken up into seven age groups where we noticed a somewhat balance number of people in each group. From the box plot, younger runners <15–25 tend to perform better, with lower runtimes, as evidenced by lower median values and a tighter interquartile range (IQR). This age group shows more variability (outliers) in performance for both genders. In the summary table, the mean runtime for this group is 11,039, with a median of 10,626. This group performs significantly better than older age groups, with a relatively smaller IQR of 2,917. The 26–35 and 36–45 age groups demonstrate the best overall marathon performance. From the box plot, both men and women in these groups have lower runtimes and reduced variability (indicated by shorter whiskers and fewer outliers). The table shows a continued decline in median runtimes as runners transition from 26–35 (median: 8,896) to 36–45 (median: 9,774). The IQR for both groups remains relatively small (~1,500), indicating more consistent performance in this age range. Men tend to outperform women in these age groups, as seen by the lower median runtimes in the blue box plots. The oldest age groups 66–75 and 76+ exhibit the largest runtimes and widest performance variability. The box plots show much larger IQRs, particularly for women, reflecting significant variation in performance. In the table, the mean and median runtimes increase for the 66–75 (mean: 16,156; median: 15,552) and 76+ (mean: 21,032; median: 20,688) groups. The IQR also expands especially for the 76+ group (IQR: 5,687), indicating a broad range of abilities within this age group.

range. The box plot demonstrates that men, even in older age groups, continue to have better performance (lower runtimes) than women. However, the performance gap between genders widens significantly after age 66, particularly in the 76+ group.

Marathon Performance Summary by Age Range

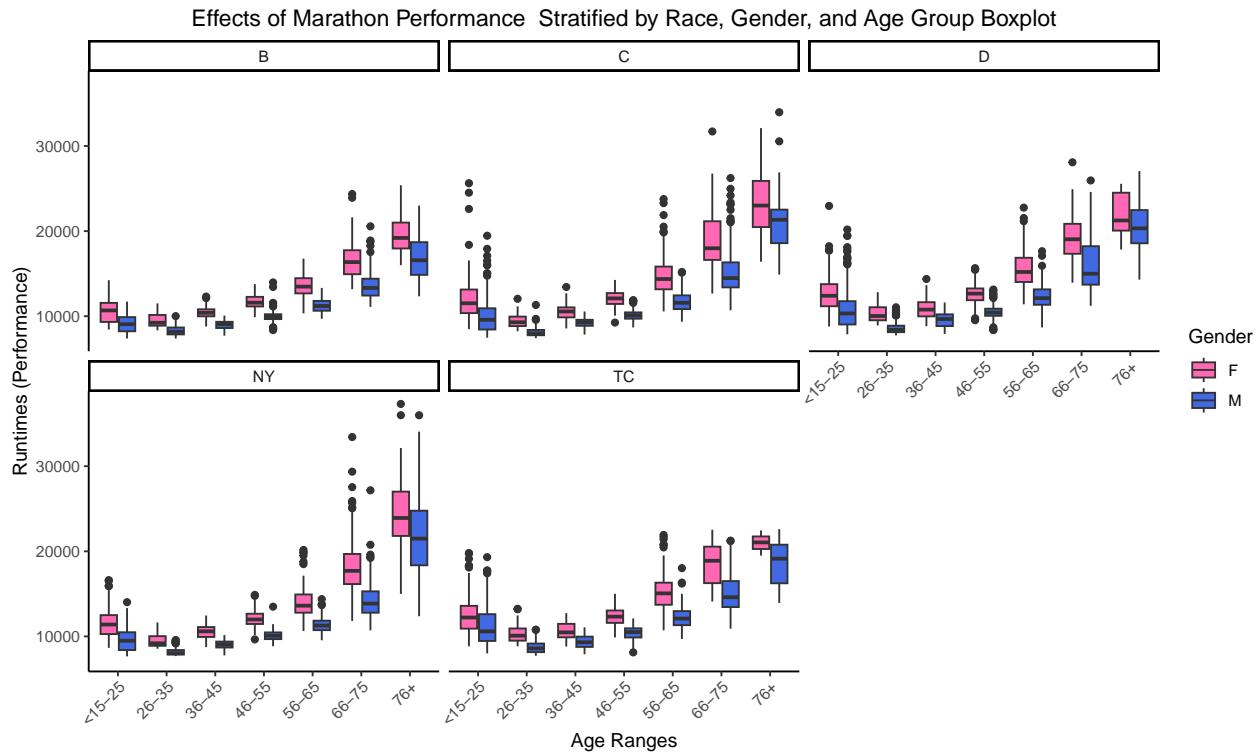
| Age Ranges | Marathon Runners | Mean Runtimes | 1Q | Median Runtimes | 3Q | IQR |
|------------|------------------|---------------|--------|-----------------|--------|-------|
| <15-25 | 1736 | 11,039 | 9,332 | 10,626 | 12,248 | 2,917 |
| 26-35 | 1840 | 9,064 | 8,194 | 8,896 | 9,684 | 1,491 |
| 36-45 | 1840 | 9,881 | 9,092 | 9,774 | 10,589 | 1,498 |
| 46-55 | 1840 | 11,174 | 10,131 | 10,968 | 12,112 | 1,981 |
| 56-65 | 1820 | 13,125 | 11,494 | 12,664 | 14,374 | 2,880 |
| 66-75 | 1463 | 16,156 | 13,700 | 15,552 | 17,964 | 4,264 |
| 76+ | 534 | 21,032 | 17,924 | 20,688 | 23,611 | 5,687 |

Effects of Age on Marathon Performance in Men and Women BoxPlot



Based on the **Effects on Marathon Performance Stratified by Race, Gender, and Age Group Boxplot**, I examined the performances of men and women by age group by each race. In the Boston Marathon, both men and women display similar runtime performance in the <15–25 age group, with a slight male advantage in the 26–35 age group that indicates low runtimes. In the case of the 56 and older groups, there is a noticeable increase in runtimes, with women showing greater variability with larger IQR representation and longer runtimes compared to men. In the Chicago Marathon, there is an indication of great performance within the 26–35 age range having the lowest runtimes. Also there is a gradual decline in performance from 36–45 and 46–55 age groups, with a noticeable increase in runtimes for both genders. For the groups 56–65 to 76, women show more variability and slower performance than men in these age groups. The Duluth Grandma's Marathon, there is an indication of an increase in runtimes for both men and women, with women signifying higher runtimes and greater variability within the 66–75 and 76+ age groups. In the

Twin Cities Marathon, runtimes increase markedly for both men and women after the age of 55. Women show higher variability and longer runtimes, particularly in the 66–75 and 76+ age groups. The New York City Marathon there is a decline in performance where the 36–45 and 46–55 groups indicates women showing slightly higher runtimes than men.



Aim 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

To begin this analysis, a linear model will be constructed to demonstrate the relationship and impact of environmental conditions on marathon performance, considering differences across age and gender. The outcome variable is the marathon runners' runtimes, analyzed in conjunction with environmental conditions, gender, and age. Upon examining the model, it was noticed that the weather parameters highlighted in blue, such as Dry Bulb Temp C ($p=0.00001$), Percent Relative Humidity ($p=0.00001$), Black Globe Temp C ($p=0.0000$), Dew Point in C ($p=0.00003$), Solar Radiation ($p=0.0000$), and Wet Bulb Globe Temp ($p=0.00001$), demonstrated statistical significance ($p<0.05$). Other significant parameters include the interaction terms, such as Age (yr) ($p=0.00000$), GenderM ($p=0.00001$), Dry Bulb Temp C:Age (yr) ($p=0.00002$), Percent Relative Humidity:Age (yr) ($p=0.0000$), Black Globe Temp C:Age (yr) ($p=0.00023$), Percent Relative Humidity:Gender ($p=0.01862$), Age (yr):GenderM ($p=0.01671$), Age (yr):Wet Bulb Globe Temp ($p=0.00017$), Solar Radiation:Age (yr) ($p=0.00000$), and Dew Point in C:Age (yr) ($p=0.00706$). These significant weather parameters were used in further analysis in the Supplementary Materials below.

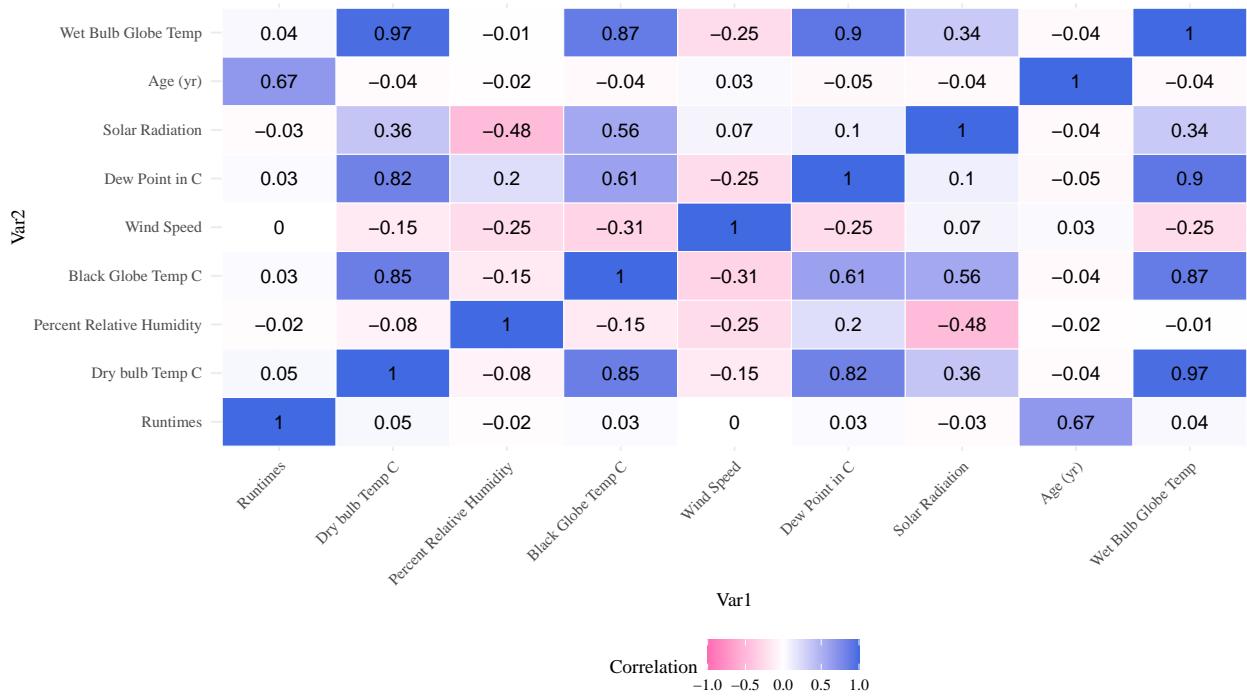
By analyzing the statistical significance of weather parameters, correlations with ages were observed. Runtimes and Age have a correlation of (0.67), indicating a strong positive relationship; as runners age, marathon runtimes tend to increase. Dry Bulb Temp C and Black Globe Temp share a strong correlation of (0.85), suggesting that these measurements, which gauge different aspects of weather temperature, tend to increase together. Wet Bulb Globe Temp and Dry Bulb Temp C exhibit a very strong correlation of (0.97), indicating that wet bulb globe temperature and dry bulb temperature often rise together. Dew Point in C and Wet Bulb Globe Temp have a strong correlation of (0.90), showing that higher dew points (indicating more

Table 3: Impact of Environmental Conditions on Marathon Performance by Age and Gender

| | Estimate | Std. Error | t value | Pr(> t) |
|--|--------------------|------------------|------------------|----------------|
| (Intercept) | 2411.77125 | 578.71285 | 4.16747 | 0.00003 |
| ‘Dry bulb Temp C‘ | -552.88575 | 122.75832 | -4.50386 | 0.00001 |
| ‘Percent Relative Humidity‘ | 21.06573 | 2.97023 | 7.09228 | 0.00000 |
| ‘Black Globe Temp C‘ | -270.50909 | 65.07994 | -4.15657 | 0.00003 |
| ‘Wind Speed‘ | -24.68197 | 21.22259 | -1.16301 | 0.24485 |
| ‘Dew Point in C‘ | -313.60541 | 92.41151 | -3.39358 | 0.00069 |
| ‘Solar Radiation‘ | 4.71062 | 0.57781 | 8.15250 | 0.00000 |
| ‘Age (yr)‘ | 221.49195 | 11.16442 | 19.83909 | 0.00000 |
| ‘Wet Bulb Globe Temp‘ | 1288.98835 | 290.41680 | 4.43841 | 0.00001 |
| GenderM | -1890.99700 | 419.11095 | -4.51192 | 0.00001 |
| ‘Dry bulb Temp C‘:‘Age (yr)‘ | 10.39954 | 2.40581 | 4.32268 | 0.00002 |
| ‘Percent Relative Humidity‘:‘Age (yr)‘ | -0.55316 | 0.05694 | -9.71409 | 0.00000 |
| ‘Black Globe Temp C‘:‘Age (yr)‘ | 4.72352 | 1.28324 | 3.68095 | 0.00023 |
| ‘Wind Speed‘:‘Age (yr)‘ | 0.31655 | 0.40853 | 0.77485 | 0.43845 |
| ‘Dew Point in C‘:‘Age (yr)‘ | 4.88779 | 1.81383 | 2.69473 | 0.00706 |
| ‘Solar Radiation‘:‘Age (yr)‘ | -0.12420 | 0.01132 | -10.97440 | 0.00000 |
| ‘Age (yr)‘:‘Wet Bulb Globe Temp‘ | -21.46104 | 5.70079 | -3.76457 | 0.00017 |
| ‘Age (yr)‘:GenderM | -6.54622 | 2.73514 | -2.39338 | 0.01671 |
| ‘Dry bulb Temp C‘:GenderM | -36.56022 | 85.61722 | -0.42702 | 0.66937 |
| ‘Percent Relative Humidity‘:GenderM | 4.82247 | 2.04912 | 2.35344 | 0.01862 |
| ‘Black Globe Temp C‘:GenderM | -24.62302 | 45.15141 | -0.54534 | 0.58553 |
| ‘Wind Speed‘:GenderM | 7.71624 | 14.53336 | 0.53093 | 0.59548 |
| ‘Dew Point in C‘:GenderM | -33.38632 | 64.05318 | -0.52123 | 0.60222 |
| ‘Solar Radiation‘:GenderM | 0.65298 | 0.39924 | 1.63557 | 0.10196 |
| ‘Wet Bulb Globe Temp‘:GenderM | 84.32974 | 202.04485 | 0.41738 | 0.67641 |

moisture in the air) tend to be associated with higher wet bulb globe temperatures, reflecting the impact of humidity on heat stress. The correlation of (0.67) between **Age** and **Runtimes** is higher than most correlations with environmental conditions, suggesting that age has a stronger impact on performance than many environmental factors. This indicates that age plays a significant role in marathon performance, regardless of environmental conditions.

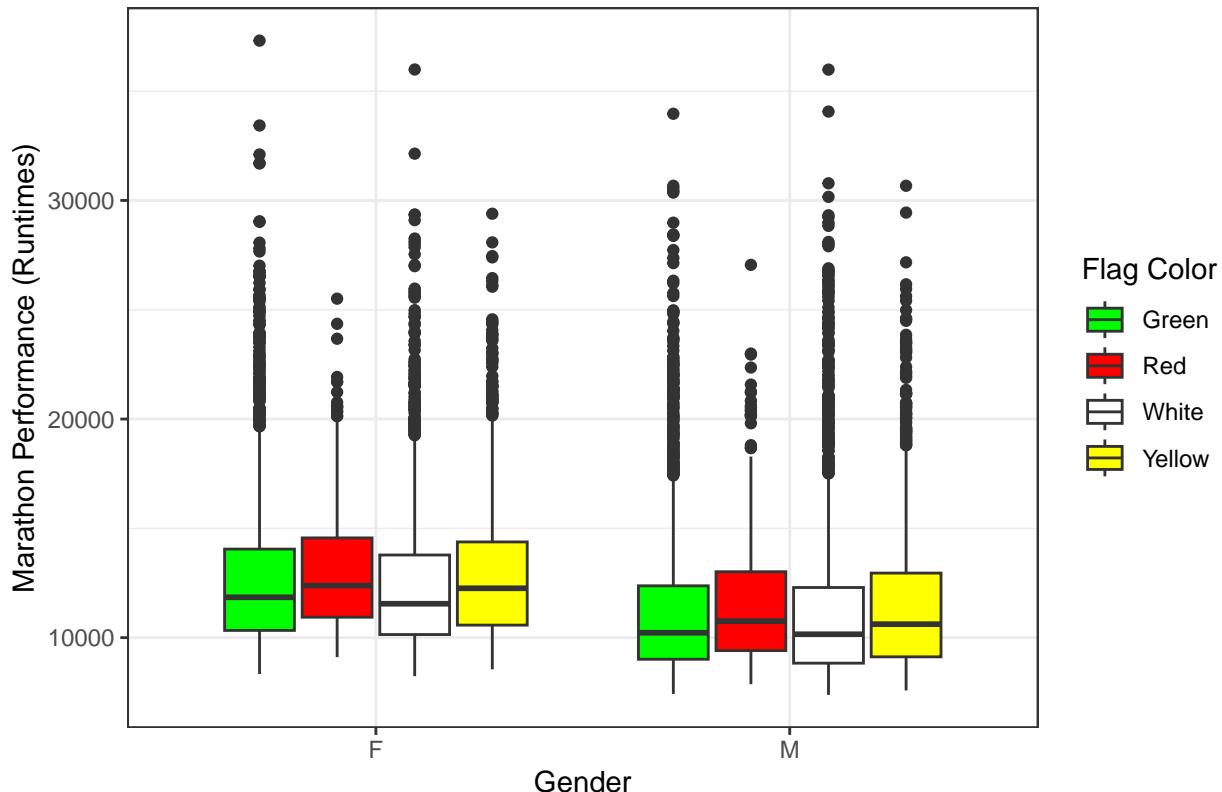
Correlation of Environmental Conditions on Marathon Performance



Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

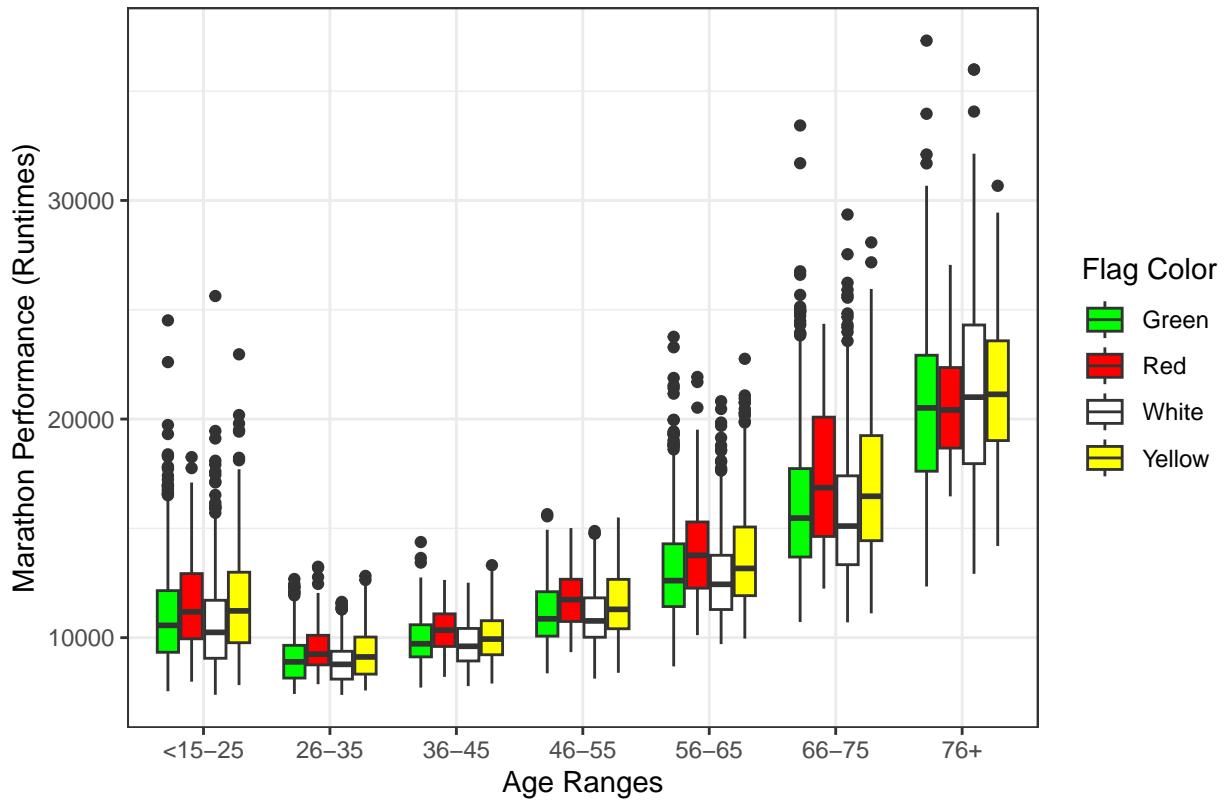
According to the **Marathon Performance by Flag Conditions Stratified by Gender** both gender runners in Green flag conditions (indicating safer weather with lower WBGT) generally have faster runtimes, as seen by the lower median performance values compared to those running under Red, White, and Yellow flag conditions. In more severe environmental conditions, such as Red and Yellow flags, which indicate high WBGT levels, marathon performance tends to be slower, with median runtimes higher than in Green flag conditions. This suggests that higher temperatures and humidity levels significantly degrade performance. The wider spread of the box plots in Red and Yellow flags indicates greater variability in performance under these conditions, likely due to varying levels of heat tolerance among runners. While both male and female runners exhibit similar trends in response to changing flag conditions, female runners appear to have slightly higher median runtimes across all flag conditions. However, the pattern of slower performance under more severe weather conditions holds for both genders.

Marathon Performance by Flag Conditions Stratified by Gender

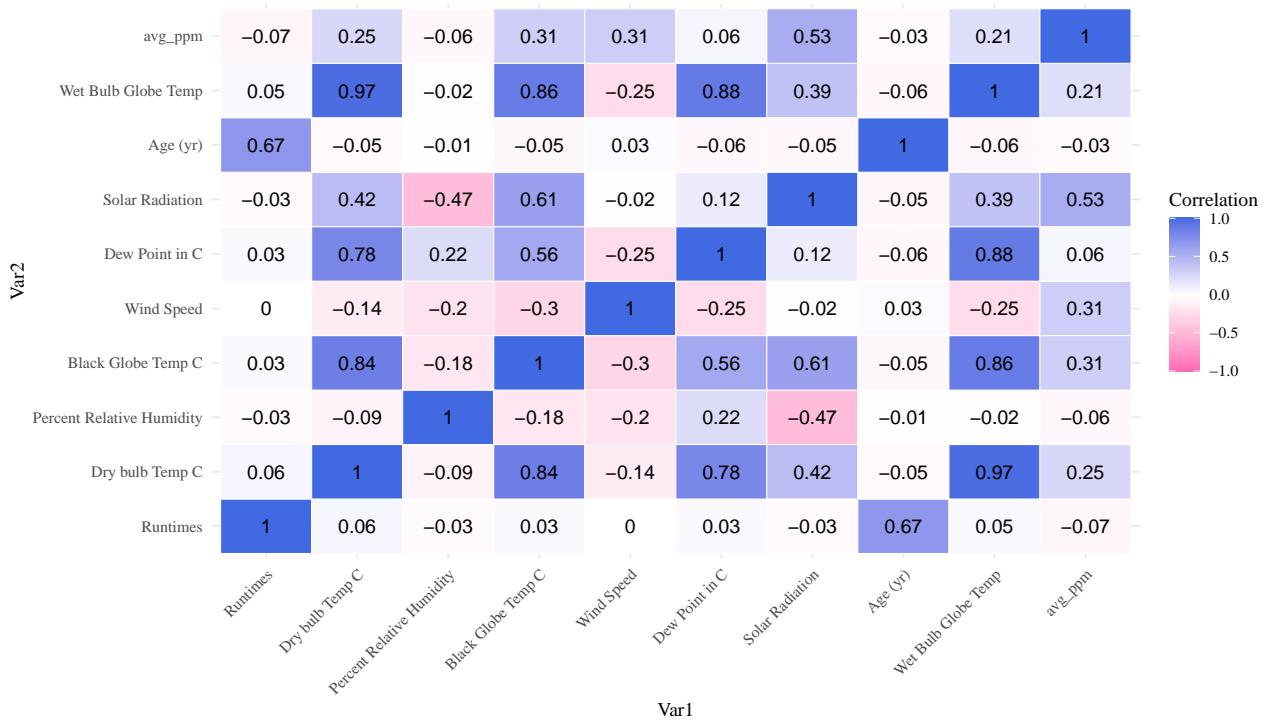


According to the **Marathon Performance by Flag Conditions Stratified by Age Ranges**, the younger age ranges like <15-25 and 26-35 categories, marathon performance is generally faster, with lower median runtimes and narrower interquartile ranges (IQR) across all flag conditions. The spread of runtimes in younger runners is relatively smaller across all flag conditions, suggesting more consistent performance regardless of environmental conditions. However, a small trend is observed where Red and Yellow flags (indicating harsher conditions) are associated with slightly slower runtimes compared to Green flags. In the 56-55 and 56-65 age groups, there is a noticeable increase in median runtimes, especially under Red and Yellow flag conditions. Runners in these middle-aged categories tend to perform more slowly in extreme heat, as indicated by the shift in the median and the broader spread of runtimes. The variability (as indicated by wider IQRs and longer whiskers) increases in the middle age groups, particularly under extreme conditions, suggesting that performance becomes more inconsistent as age increases and environmental conditions worsen.

Marathon Performance by Flag Conditions Stratified by Age Ranges



Weather Parameters with the Largest Impact on Marathon Performance Correlation Plot



In the Weather Parameters with the Largest Impact on Marathon Performance Correlation Plot, Dry Bulb Temp C and Black Globe Temp C have a strong positive correlation of (0.84). This makes sense

as both variables represent temperature-related measurements. Also the Wet Bulb Globe Temp and Dry Bulb Temp C are also highly correlated (0.97), indicating that as dry bulb temperature increases, wet bulb temperature and black globe temperature tend to increase. The correlation between Dew Point in C and Wet Bulb Globe Temp (0.88) is also strong, reflecting that higher dew points (indicating more moisture in the air) typically occur alongside increased wet bulb temperatures, suggesting a higher overall thermal burden. Age shows a moderate positive correlation (0.67) with Runtimes, indicating that as runners get older, their marathon runtimes tend to increase because older runners take longer to complete the marathon. This is an expected result, as endurance performance often declines with age. Based on the correlation plot, Wind Speed has a very low positive correlation at (0.03) on Age (yr) indicating that wind speed has a minimal influence on age-related variations in marathon performance. (Check Supplementary Materials for extra analysis).

Conclusion and Discussion

In conclusion, the data examining marathon performance across a broad age range from 14 to 91 years reveals a clear trend: runtimes decrease, indicating improved performance, from mid twenties to about 45 years of age, peaking typically in the mid-twenties. This suggests that younger athletes, especially those in their mid-twenties or younger, tend to perform better compared to their older counterparts. Beyond this peak, there is a consistent increase in runtimes with advancing age, signaling a decline in performance. This decline is more pronounced in women, particularly noticeable in the steeper performance drop observed around the ages of 60-65. Men, while also experiencing a decline, consistently post faster runtimes across all ages. The gap between the genders widens with age, with men maintaining faster runtimes compared to women in older age groups. This widening gap underscores a more significant impact of aging on female marathon runners, highlighting a greater decline in performance as they age.

The exploratory data analysis demonstrates considerable impact of environmental conditions on marathon performance across both genders and age groups. According to *Aerobic Performance is Degraded Despite Modest Hypothermia in Hot Environments*, "Endurance exercise performance is degraded within increasing environmental temperature, and the decline in performance is associated with warmer temperature is magnified with longer-distance events such as the marathon foot race (Ely, B. R., et.al). Studies have shown that marathon runners prefer cooler temperatures rather than warmer and humid temperatures due to the body regulating excessive amounts of heat. Running in cooler weather conditions, can reduce risks of heat exhaustion or heat strokes which are common in warmer conditions. When the temperature is lower, the cardiovascular system doesn't have to work as hard to cool the body .This help runners maintain pace without their heart rate increasing as much in hotter conditions. Runners typically achieve better performance under Green flag conditions, which denote safer weather scenarios with lower Wet Bulb Globe Temperature (WBGT), evidenced by the lower median runtimes in comparison to more severe conditions marked by Red, White, and Yellow flags. With the transition to Red and Yellow flags, indicative of higher WBGT levels, there is a distinct drop in performance, seen in the increased median runtimes. This performance decline under harsher weather conditions can be linked to elevated temperatures and humidity, which are characteristic of these flags. Furthermore, the increased variability in performance under Red and Yellow flags may suggest a broad spectrum of individual heat tolerances among runners. Although both male and female runners exhibit similar patterns in response to changing flag conditions, females consistently display slightly higher median runtimes in all conditions indicating slower runtimes.

Supplementary Material

Marathon Performance vs. Dry Bulb Temperature by Age Ranges

For all age groups, the effect of dry bulb temperature (ambient air temperature) on marathon performance appears to be relatively small. The scatterplots show only a slight upward trend in runtimes as dry bulb temperature increases, indicating that higher temperatures may have a mild negative impact on performance. However, the slope of the black trend lines is nearly flat for most age groups, suggesting that temperature changes in the given range (roughly 10°C to 25°C) do not strongly affect marathon runtimes.

Men and women across age groups show similar patterns in how dry bulb temperature affects performance. While women generally exhibit slightly higher runtimes than men within each age group, the impact of temperature on performance does not appear to differ significantly between the genders. This suggests that temperature influences men and women similarly, with no large gender-based differences in how environmental heat affects marathon performance.

As age increases, there is a noticeable rise in median runtimes, particularly in the older age groups 56–65, 66–75, and 76+. While the influence of temperature is small across all groups, the general increase in runtimes with age remains evident, independent of temperature. For the 76+ group, runtimes are much higher overall, but again, the effect of temperature is minimal. This suggests that age-related declines in performance are more pronounced than the impact of environmental conditions like temperature, especially in older runners.

Marathon Performance vs. Relative Humidity Stratified by Age Ranges

Amongst all the age ranges, the effect of relative humidity on marathon performance is minimal, as indicated by the nearly flat trend lines across the scatterplots. This suggests that variations in relative humidity, ranging from 0% to 100%, have a limited impact on marathon runtimes for both men and women. This trend is consistent across all age groups, indicating that relative humidity alone does not significantly affect performance. In the **Marathon Performance vs. Relative Humidity Stratified by Age Ranges** men have lower runtimes than women across all age groups, but the impact of relative humidity on performance appears similar for both genders. The black trend lines, representing the overall relationship between humidity and performance, remain flat for both men and women, suggesting that both genders are equally unaffected by changes in humidity.

Marathon Performance vs. Black Globe Temperature Stratified by Age Groups

Across most age groups, there is a slight upward trend in runtimes as Black Globe Temperature increases, especially in the 46–55, 56–65, and 66–75 age ranges. This suggests that higher Black Globe Temperatures, which represent heat stress caused by solar radiation and air temperature, may have a mild negative impact on marathon performance. Runners in these age groups experience slightly longer runtimes at higher temperatures, indicating that environmental heat does begin to affect performance, although the effect is not very strong.

Marathon Performance vs. Wet Bulb Globe Temp Stratified by Age Groups

The trend lines in all age groups show an upward slope that indicates a small negative impact of Wet Bulb Globe Temperature (WBGT) on marathon performance. As WBGT increases (which measures both temperature and humidity), runtimes tend to increase slightly, but the overall effect is minimal. The trend lines for both men and women show similar trajectories, indicating that WBGT affects both genders equally, with no significant gender-based difference in how environmental heat and humidity influence performance. The performance gap between men and women remains consistent regardless of WBGT levels.

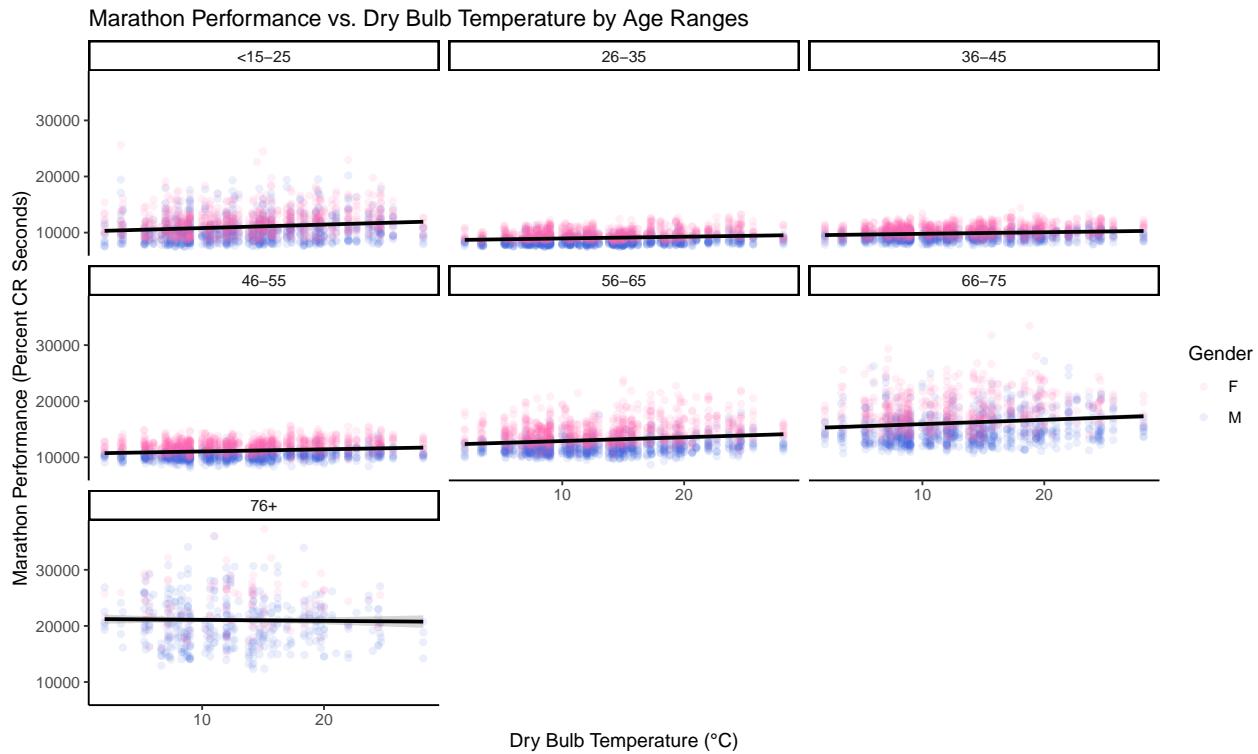
Marathon Performance vs. Solar Radiation Stratified by Age Groups

Across all age groups, there is little to no effect of solar radiation on marathon performance. The trend lines across different levels of solar radiation (ranging from 0 to 750 units) are mostly flat, indicating that

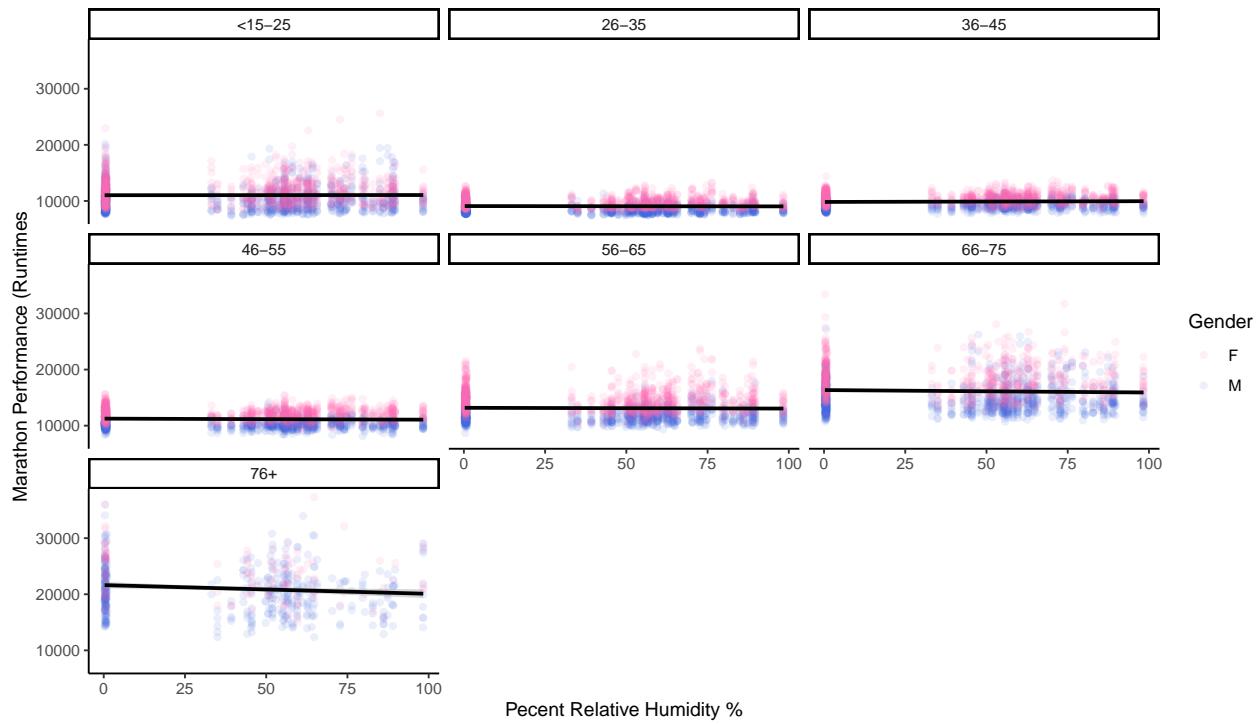
increased solar radiation does not have a strong impact on runtimes for either gender. This indicates that solar exposure alone does not significantly affect marathon performance.

Marathon Performance vs. Black Globe Temperature Stratified by Age Groups

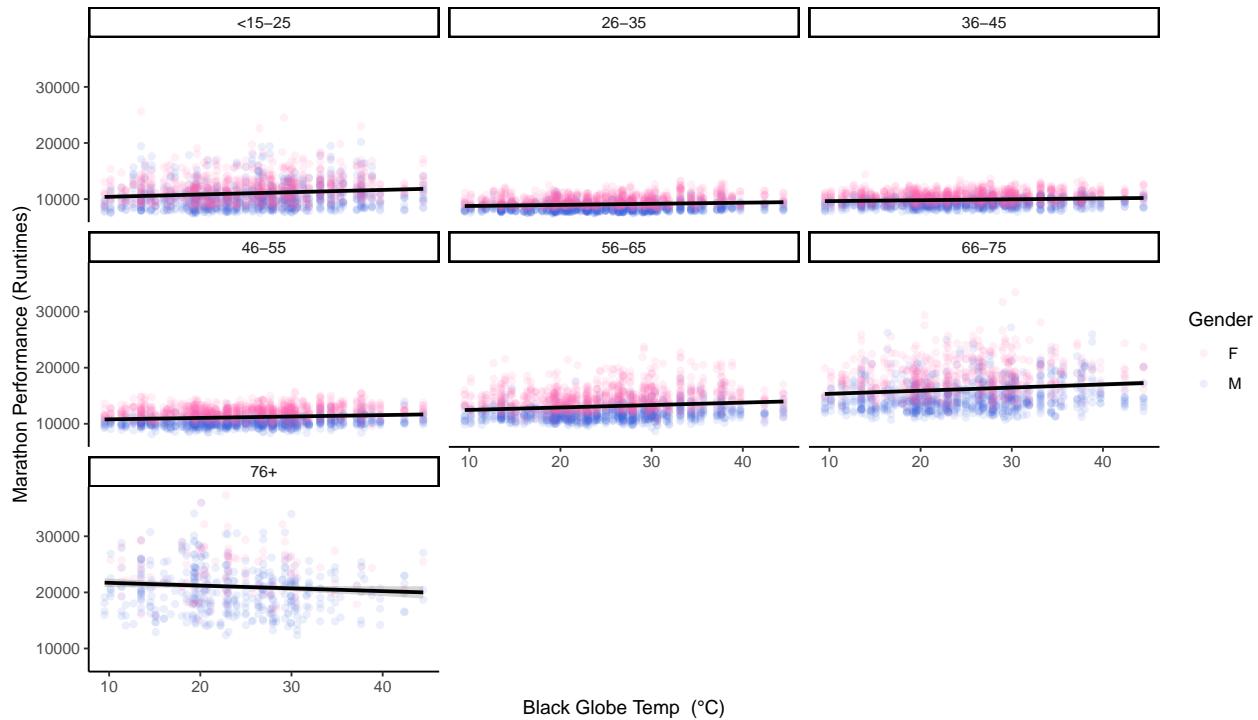
The slope of the trend lines across all age groups is generally flat, suggesting that Black Globe Temperature (BGT) does not have a strong impact on marathon performance, especially in younger age groups (<15-25, 26-35, 36-45). This indicates that younger runners are more resilient to changes in environmental heat, as their performance remains fairly consistent across different temperature ranges.



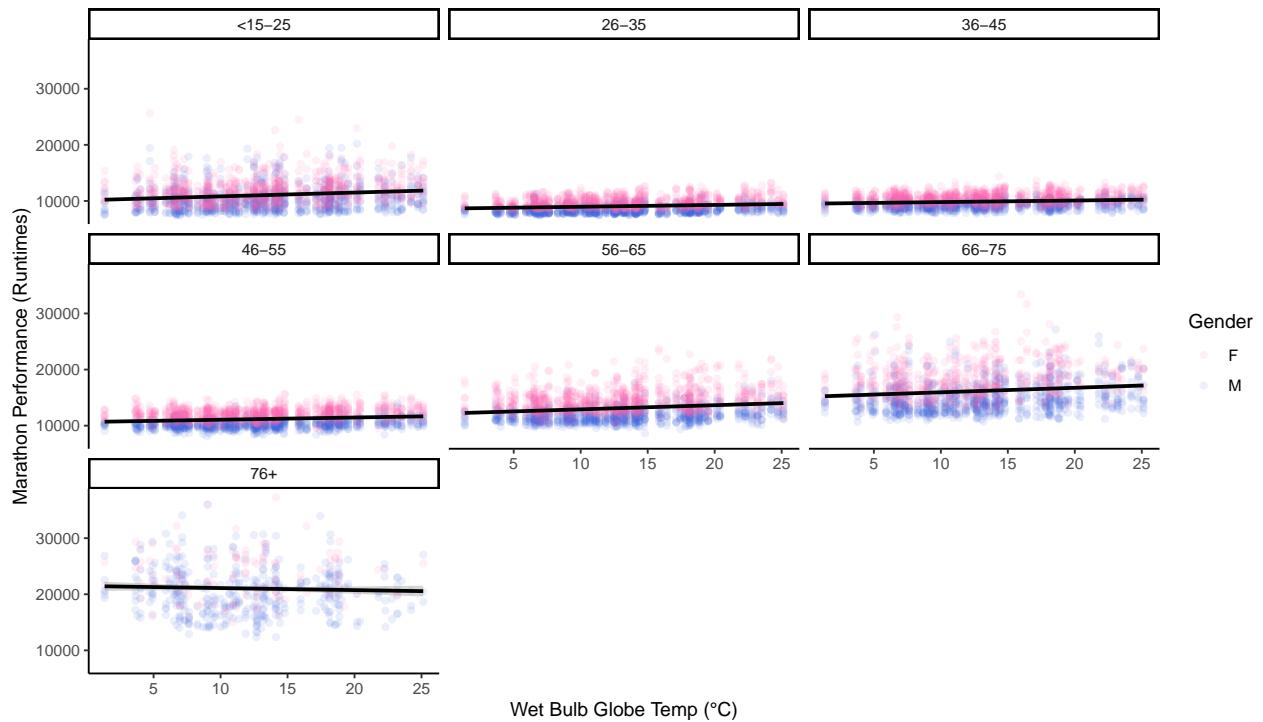
Marathon Performance vs. Relative Humidity Stratified by Age Ranges



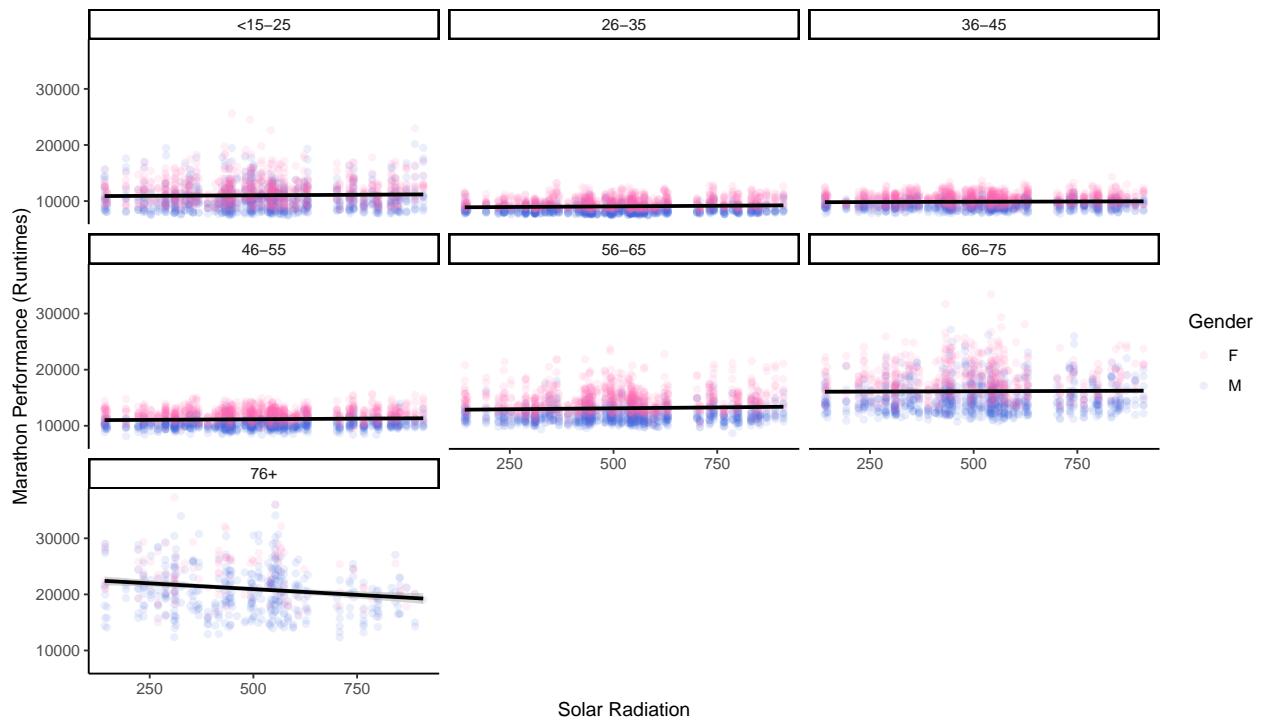
Marathon Performance vs. Black Globe Temperature Stratified by Age Groups



Marathon Performance vs. Wet Bulb Globe Temp Stratified by Age Groups



Marathon Performance vs. Solar Radiation Stratified by Age Groups



Marathon Performance vs. Black Globe Temperature Stratified by Age Groups

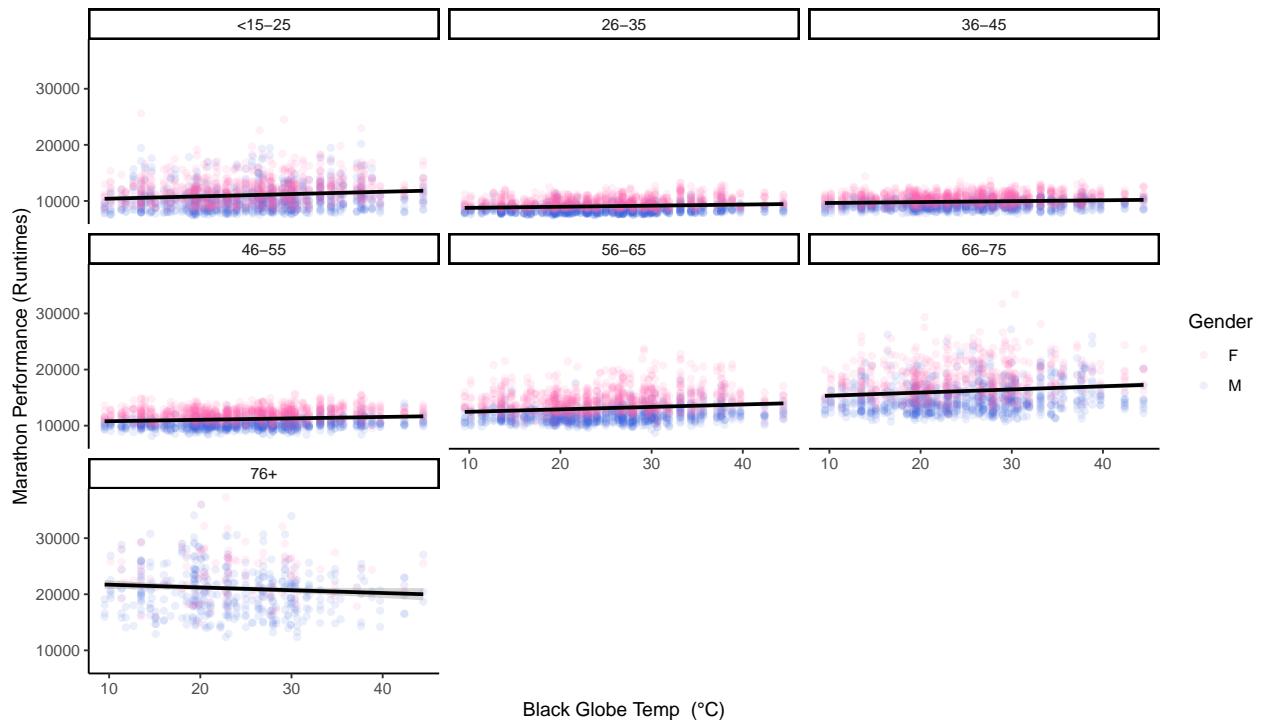


Table 4: Marathon Runners' Data Description

| Variables | Missing Data | Type | Description |
|---------------------------|--------------|-------------|--|
| Age (yr) | 0 | Numeric | Age (yr) represents the ages of the participants. |
| Black Globe Temp C | 491 | Numeric | Black Globe Temp Celcius indicates how hot it feels in direct sunlight. It considers temperature, humidity, wind speed, sun angle, and cloud cover to provide a holistic view of the stress placed on the body in hot environments. |
| CR | 0 | HMS/Numeric | CR is the course record for each marathon. |
| Dew Point in C | 491 | Numeric | Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form. If the air were to be cooled even more, water vapor would have to come out of the atmosphere in the liquid form, usually as fog or precipitation. |
| Dry bulb Temp C | 491 | Numeric | Dry bulb Temp Celcius is the air temperature without taking into account of the humidity or any moisture. |
| Flag | 491 | Character | Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C, Red= WBGT >23-28C, and Black= WBGT > 28C |
| Gender | 0 | Character | Gender is represented by F= Female and M= Male. |
| Percent CR | 0 | Numeric | Percent CR is the percent off current course record for gender. |
| Percent Relative Humidity | 491 | Numeric | Percent Relative Humidity how much moisture is in the air compared to the maximum amount of moisture the air can hold at a given temperature. Gives an idea of how humid it feels outside. |
| Race | 0 | Character | Race represents the marathons the participants competed, including the B=Boston Marathon, C= Chicago Marathon, NY= New York City Marathon, T= Twin Cities Marathon (Minneapolis,MN), D= Grandma's Marathon (Duluth, MN). |
| Race_Seconds | 0 | Numeric | Race_Seconds is the course record measured in seconds. |
| Runtimes | 0 | Numeric | Runtimes is the converted gender percentage into seconds. |
| Solar Radiation | 491 | Numeric | Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels through space and reaches the Earth as light and heat. |
| Wet Bulb Globe Temp | 491 | Numeric | Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar radiation on humans. Formula $WBGT = 0.7 \times Tw + 0.2 \times Tg + 0.1 \times Td$. |
| Wet bulb Temp C | 491 | Numeric | Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity in the air. Wet bulb temperature gives you an idea of how temperature feels when you take humidity into account. |
| Wind Speed | 491 | Numeric | Wind Speed in Km/hr. |
| Year | 0 | Numeric | Years represented in the dataset ranging from 1993-2016. |

References

- Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. *Med Sci Sports Exerc*, 42(1), 135-41.
- Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. *Medicine and science in sports and exercise*, 39(3), 487-493.
- Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. *Journal of applied physiology*, 95(6), 2598-2603.
- Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., ... & Millet, G. Y. (2022). Sex differences in endurance running. *Sports medicine*, 52(6), 1235-1257.
- Yanovich, R., Ketko, I., & Charkoudian, N. (2020). Sex differences in human thermoregulation: relevance for 2020 and beyond. *Physiology*, 35(3), 177-184.

```

# Code Appendix

knitr::opts_chunk$set(warning = FALSE,
                      message = FALSE,
                      echo = FALSE,
                      fig.align = "center")

#Data
library(readr)
marathon_dates <- read_csv("marathon_dates.csv")
course_record <- read_csv("course_record.csv")
aqi_values <- read_csv("aqi_values.csv")
project1 <- read_csv("project1.csv")
#Packages
library(lubridate)
library(tidyverse)
library(tidyr)
library(dplyr)
library(naniar)
library(visdat)
library(kableExtra)
library(knitr)
library(gridExtra)
library(ggridges)
library(gt)
library(ggwordcloud)
library(ggplot2)
library(magick)
library(ggplot2)
library(corrplot)
library(reshape2)

# Path to your image
fig_path <- "/Users/diahminhawkins/Documents/GitHub/Project1/weather.png"

# Load the image using magick
img <- image_read(fig_path)

# Convert image to raster for use in ggplot
img_raster <- as.raster(img)

# Example data
words <- c("Boston", "New York City", "Minneapolis", "Grandma's", "Chicago",
          "Race", "Age", "Gender", "Weather", "Performance",
          "Wet Bulb Globe Temperature", "Humidity", "Black Globe Temp C", "Dew Point in C", "Dry bulb",
          "Percent Relative Humidity", "Solar Radiation", "Wet Bulb Globe Temp",
          "Wet Bulb Temp C", "Wind Speed")

```

```

frequencies <- c(1, 18, 1, 16, 14, 55, 5, 1, 4, 42, 3, 14, 14, 18, 16, 4, 7, 9, 10, 22, 55)

new_frame <- data.frame(words, frequencies)

# Generate the word cloud on top of the image background
ggplot(new_frame, aes(label = words, size = frequencies)) +
  # Add the image background
  annotation_raster(img_raster, xmin = -Inf, xmax = Inf, ymin = -Inf, ymax = Inf) +

  # Generate the word cloud
  geom_text_wordcloud(aes(color = frequencies)) +
  scale_size_area(max_size = 10) +

  # Customize the colors of the words
  scale_color_gradient(low = "yellow", high = "red") +

  # Remove axis titles and labels since we want the word cloud only
  theme_void()

# Currently, there has been an increase in marathon participation and performance in the past two decades
#Course Record Data Management
course_record<- course_record%>%
  mutate(Race_Seconds= as.numeric(hms(course_record$CR)))

#Change column names from Sex... to Gender to match project1 dataset
colnames(course_record)[colnames(course_record) == "Sex (0=F, 1=M)"] ="Gender"

# Change the Race variable to a character variable
course_record <- course_record %>%
  mutate(Race = as.character(Race))

# Project 1 Data Management
#Change colnames for a more readable and understanding approach
colnames(project1)[colnames(project1) == "Sex (0=F, 1=M)"] ="Gender"
colnames(project1)[colnames(project1) == "Race (0=Boston, 1=Chicago, 2=NYC, 3=TC, 4=D)"] ="Race"
colnames(project1)[colnames(project1) == "Td, C"] ="Dry bulb Temp C"
colnames(project1)[colnames(project1) == "Tw, C"] ="Wet bulb Temp C"
colnames(project1)[colnames(project1) == "%rh"] ="Percent Relative Humidity"
colnames(project1)[colnames(project1) == "Tg, C"] ="Black Globe Temp C"
colnames(project1)[colnames(project1) == "SR W/m2"] ="Solar Radiation"
colnames(project1)[colnames(project1) == "DP"] ="Dew Point in C"
colnames(project1)[colnames(project1) == "Wind"] ="Wind Speed"
colnames(project1)[colnames(project1) == "WBGT"] ="Wet Bulb Globe Temp"
colnames(project1)[colnames(project1) == "%CR"] ="Percent CR"
#change gender if 0 to F to represent female
project1$Gender<-ifelse(project1$Gender== "0","F","M") #change gender if 0 to F to represent female else M

# Mutate the Race names from numbers to the marathon cities for better understanding
project1 <- project1 %>%

```

```

mutate(Race = case_when(
  Race == "0" ~ "B",
  Race == "1" ~ "C",
  Race == "2" ~ "NY",
  Race == "3" ~ "TC",
  Race == "4" ~ "D"))

#Change the Race variable to a character variable
project1 <- project1 %>%
  mutate(Race = as.character(Race))

# Merge the two dataframes
course_record_project1<-left_join(project1,course_record,
                                     by= c("Race", "Gender", "Year"))

# Change the Course Percentage %CR into course seconds
course_record_project1 <- course_record_project1 %>%
  mutate(Runtimes = Race_Seconds * (1 + (^Percent CR` / 100)))

#Calculate the sum of all the missing data (NAs)
sum_of_na<-sum(is.na(course_record_project1))
#Examine the data
course_record_project1%>% vis_dat()
vis_miss(course_record_project1)
course_record_project1%>% glimpse()

#Get all the missing data from each column
Missing_Data<- sapply(course_record_project1, function(x) sum(is.na(x)))
# Convert to dataframe
Missing_Data_df <- data.frame(ColumnName = names(Missing_Data), `Missing Data` = Missing_Data)

# Set names for the dataframe columns if necessary
names(Missing_Data_df) <- c("Variables", "Missing Data")

# Calculate the total number of rows in the dataset
total_rows <- nrow(course_record_project1)

#Create Missing Data Summary
missing_data_summary <-Missing_Data_df %>%
  filter(Variables %in% c('Black Globe Temp C', 'Dew Point in C', 'Dry bulb Temp C', 'Flag',
                         'Percent Relative Humidity', 'Solar Radiation', 'Wet Bulb Globe Temp',
                         'Wet bulb Temp C', 'Wind Speed')) %>%
  mutate(Percent_Missing = (^Missing Data` / total_rows) * 100) %>%
  select(Variables, `Missing Data`, Percent_Missing)

# Convert to a gtsummary table
missing_data_summary %>%
  gt() %>%
  tab_header(
    title = "Missing Data Summary for Environmental Variables"
  ) %>%
  cols_label(

```

```

Variables = "Variables",
`Missing Data` = "Missing Values",
Percent_Missing = "Percentage Missing (%)"
) %>%
fmt_number(
  columns = vars(Percent_Missing),
  decimals = 2 # Format percentage to two decimal places
)

#Create variable table dataframe with description of the Marathon Dat
Variables_table<- data_frame(
  Variables= c("Race", "Year", "Gender", "Flag", "Age (yr)",
  "Percent CR", "Dry bulb Temp C","Wet bulb Temp C",
  "Percent Relative Humidity", "Black Globe Temp C","Solar Radiation",
  "Dew Point in C", "Wind Speed" , "Wet Bulb Globe Temp","CR",
  "Race_Seconds","Runtimes"),
  Type= c("Character", "Numeric", "Character", "Character","Numeric",
  "Numeric","Numeric", "Numeric", "Numeric","Numeric","Numeric",
  "Numeric","Numeric", "HMS/Numeric", "Numeric","Numeric"),
  Description= c("Race represents the marathons the participants competed, including the B=Boston Marathon,
  C= Chicago Marathon, NY= New York City Marathon,T= Twin Cities Marathon (Minneapolis,MN),
  D= Grandma's Marathon (Duluth, MN).",
  "Years represented in the dataset ranging from 1993-2016.",
  "Gender is represented by F= Female and M= Male.",
  "Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C,
  Red= WBGT >23-28C, and Black= WBGT > 28C",
  "Age (yr) represents the ages of the participants.",
  "Percent CR is the percent off current course record for gender.",
  "Dry bulb Temp Celcius is the air temperature without taking into account of the humidity and moisture.",
  "Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity in the air.",
  "Percent Relative Humidity how much moisture is in the air compared to the maximum amount of moisture it can hold at a given temperature.",
  "Black Globe Temp Celcius indicates how hot it feels in direct sunlight. It considers both the direct and reflected solar radiation.",
  "Solar Radiation in Watts per meter squared is the energy emitted by the sun, which drives the temperature of the earth and its atmosphere.",
  "Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure) for the air to become saturated with water vapor at a given temperature and pressure.",
  "Wind Speed in Km/hr.",
  "Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed and solar radiation on the perceived temperature of the air.",
  "CR is the course record for each marathon.",
  "Race_Seconds is the course record measured in seconds.",
  "Runtimes is the converted gender percentage into seconds."
)
)

Missing_Data_df$Variables <- as.character(Missing_Data_df$Variables)
Variables_table$Variables <- as.character(Variables_table$Variables)
merged_df <- merge(Missing_Data_df, Variables_table, by = "Variables", all = TRUE)

# Create the table with kable and customize with kableExtra
table_summary<- kable(merged_df, "latex", booktabs = TRUE, caption = "Marathon Runners' Data Description")

```

```

kable_styling(latex_options = c("striped", "scale_down")) %>%
column_spec(1, width = "3cm") %>%
column_spec(2, width = "2cm") %>%
column_spec(3, width = "2cm") %>%
column_spec(4, width = "8cm")

#Remove all na's
course_record_project1<-na.omit(course_record_project1)

#Find the minimum age
minimum_age<-min(course_record_project1$`Age (yr)`)

#Find the maximum age
maximum_age<-max(course_record_project1$`Age (yr)`)

#Create Age Ranges/ Age Breaks to Categorize the groups
course_record_project1$age_ranges<- cut(
  course_record_project1$`Age (yr)`,
  breaks = c(0, 25, 35, 45, 55, 65, 75, Inf), # Custom breaks for the new age ranges
  labels = c("<15-25", "26-35", "36-45", "46-55", "56-65", "66-75", "76+"))

)

#Get Counts By Age Group to see balance
age_range_counts <- course_record_project1 %>%
  group_by(age_ranges)%>%
  summarise(count=n())

#Marathon Performance by age by Race
marathon_performance_by_age<- course_record_project1%>%
  select(Race,Year, Gender, age_ranges, Runtimes, `Age (yr)`)%>%
  group_by(Race, Gender, age_ranges)

#Get the best course_record from each race
best_course_race<- course_record_project1 %>%
  filter(Runtimes <= Race_Seconds)%>%
  group_by(Race, Gender, age_ranges)%>%
  summarise(count=n())

# Rename some of the column names
best_course_race <- best_course_race %>%
  rename(
    `Marathon` = Race,           # Rename 'Race' to 'Race Name'
    `Gender` = Gender,          # Keep the 'Gender' column as is (optional)

```

```

`Age Range` = age_ranges,      # Rename 'age_ranges' to 'Age Range'
`Number of Participants` = count # Rename 'count' to 'Number of Participants'
)
# Create the table with the new column names and specified styling
best_course_race %>%
  kbl(caption = "<div style='text-align:center; font-size:24px; font-weight:bold;'>Marathon Runners with the Best Course Record</div>",
  kable_classic(full_width = F, html_font = "Cambria", font_size = 20) %>%
  kable_styling(position = "center", font_size = 16)

worst_course_race <- course_record_project1 %>%
  filter(Runtimes >= Race_Seconds) %>%
  group_by(Race, Gender, age_ranges)%>%
  summarise(count=n())

just_gender_age_ranges<- course_record_project1 %>%
  filter(Runtimes >= Race_Seconds) %>%
  group_by(Gender, age_ranges)%>%
  summarise(count=n())

worst_course_race %>%
  kbl(caption = "Number of Marathon Runners that Did not beat the Course Record by Race, Gender, and Age",
  kable_classic(full_width = F, html_font = "Times New Roman", font_size= 20)

# Create bar plot od the Worst Course Race varaiable for easier read
ggplot(worst_course_race, aes(x = Race, y = count, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") + # Create a bar plot, grouped by Gender
  labs(title = "Number of Marathon Runners that Did not beat the Course Record by Race, Gender, and Age",
       x = "Race",
       y = "Marathon Runners") +
  scale_fill_manual(values = c("F" = "hotpink", "M" = "royalblue")) + # colors pink for F and blue for M
  facet_wrap(~ age_ranges) +
  theme_minimal()

# Line plot of the Marathon Runners Performance by Age
age_plot<-ggplot(marathon_performance_by_age, aes(x = `Age (yr)`, y = Runtimes, color = Gender)) +
  geom_point(alpha = 0.05) +
  geom_smooth(se = FALSE, linewidth = 1.5) +
  labs(title = "Effects of Age on Marathon Performance in Men and Women",
       x = "Ages",
       y = " Runtimes (Performance)") +
  scale_color_manual(values = c("F" = "hotpink", "M" = "royalblue")) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5) )

age_plot

```

```

#Create summary statistics table
marathon_performance_summary_table <-marathon_performance_by_age%>%
  group_by(age_ranges) %>%
  summarise(
    Count=n(),
    Mean= mean(Runtimes),
    Q1_Runtime = quantile(Runtimes, 0.25, na.rm = TRUE),
    Median_Runtime = median(Runtimes, na.rm = TRUE),          # Median of Runtimes
    # Lower quartile (25th percentile)
    Q3_Runtime = quantile(Runtimes, 0.75, na.rm = TRUE),      # Upper quartile (75th percentile)
    IQR_Runtime = IQR(Runtimes, na.rm = TRUE) # IQR of Runtimes
  )

marathon_performance_summary_table %>%
  gt() %>%
  tab_header(
    title = "Marathon Performance Summary by Age Range"
  ) %>%
  cols_label(
    age_ranges = "Age Ranges",
    Count = "Marathon Runners",
    Mean = "Mean Runtimes",
    Q1_Runtime = "1Q",
    Median_Runtime = "Median Runtimes",
    Q3_Runtime = "3Q",
    IQR_Runtime = "IQR"
  ) %>%
  fmt_number(
    columns = vars(Mean, Median_Runtime, Q1_Runtime, Q3_Runtime, IQR_Runtime),
    decimals = 0 # Set decimal places for summary statistics
  )
)

# Create the boxplot to visualize the grouping
age_boxplot<-ggplot(marathon_performance_by_age, aes(x = age_ranges, y =Runtimes, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Effects of Age on Marathon Performance in Men and Women BoxPlot",
       x = "Age Ranges",
       y = "Runtimes (Performance)") +
  scale_fill_manual(values = c("F" = "hotpink", "M" = "royalblue"))+
  theme_classic()+
  theme(
    plot.title = element_text(hjust = 0.5))

age_boxplot

# Create boxplot stratified by different Races
ggplot(marathon_performance_by_age, aes(x = age_ranges, y = Runtimes, fill = Gender)) +

```

```

geom_boxplot() +
facet_wrap(~ Race) + # Facet by Race to create separate plots for each race
labs(title = "Effects of Marathon Performance Stratified by Race, Gender, and Age Group Boxplot",
x = "Age Ranges",
y = "Runtimes (Performance)") +
theme_classic()+
scale_fill_manual(values = c("F" = "hotpink", "M" = "royalblue"))+
theme(axis.text.x= element_text(angle =45, vjust= 1, hjust = 1),
plot.title = element_text(hjust = 0.5) )

# Create Dataframe to allocate for other environmental conditions
environmental_conditions <- course_record_project1 %>%
select(Race, Gender, `Age (yr)`, Runtimes, `Dry bulb Temp C`, `Wet bulb Temp C`,
`Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`, `Dew Point in C`,
`Wind Speed`, `Wet Bulb Globe Temp`, age_ranges) %>%
group_by( Gender, `Age (yr)`)

#Observing the linear model approach to see the impact on performance due to weather conditions and
# observe the difference in gender and age
# Linear to evaluate statistical significance
model <- lm(Runtimes ~ `Dry bulb Temp C` + `Percent Relative Humidity` + `Black Globe Temp C` +
`Wind Speed` + `Dew Point in C` + `Solar Radiation` + `Age (yr)` + `Wet Bulb Globe Temp` + Gender +
`Dry bulb Temp C`:`Age (yr)` + `Percent Relative Humidity`:`Age (yr)` +
`Black Globe Temp C`:`Age (yr)` + `Wind Speed`:`Age (yr)` + `Dew Point in C`:`Age (yr)` +
`Solar Radiation`:`Age (yr)` + `Wet Bulb Globe Temp`:`Age (yr)` + Gender:`Age (yr)` +
`Dry bulb Temp C`:Gender + `Percent Relative Humidity`:Gender + `Black Globe Temp C`:Gender +
`Wind Speed`:Gender + `Dew Point in C`:Gender + `Solar Radiation`:Gender +
`Wet Bulb Globe Temp`:Gender, data = environmental_conditions)

model_table <- round(summary(model)$coefficients, 5)

# Use kable to create the summary table
model_table %>%
kbl(caption = "Impact of Environmental Conditions on Marathon Performance by Age and Gender",
booktabs = TRUE, escape = FALSE, align = "c") %>%
kable_styling(full_width = FALSE, latex_options = c('hold_position'))%>%
column_spec(1, bold = TRUE, color="black", border_right = TRUE)%>%
row_spec(1, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(2, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(3, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(4, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(6, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(7, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(8, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(9, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(10, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(11, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(12, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(13, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(15, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(16, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(17, bold = TRUE, color="black", background = "#87CEEB")%>%

```

```

row_spec(18, bold = TRUE, color="black", background = "#87CEEB")%>%
row_spec(20, bold = TRUE, color="black", background = "#87CEEB")

#Create Correlation plot
# Observing and Including all the numeric variables
numeric_data <- course_record_project1%>%
  select(Runtimes, `Dry bulb Temp C` , `Percent Relative Humidity` , `Black Globe Temp C` ,
         `Wind Speed` , `Dew Point in C` , `Solar Radiation` , `Age (yr)` , `Wet Bulb Globe Temp`)

# Compute the correlation matrix using complete observations
cor_matrix <- cor(numeric_data, use = "complete.obs") # Use complete.obs to ignore NAs

# Melt the correlation matrix for ggplot2
cor_data <- melt(cor_matrix)

#Environmental conditions correlation plot
environmental_conditions_plot<-ggplot(data = cor_data, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  scale_fill_gradient2(low = "hotpink", high = "royalblue", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  labs(title = "Correlation of Environmental Conditions on Marathon Performance") +
  theme_minimal(base_family = "Times") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        legend.position = "bottom",
        plot.title = element_text(hjust = 0.5, size= 20) )

environmental_conditions_plot

#Data Management of the marathon data set and air quality
#Change Race Names
marathon_dates <- marathon_dates %>%
  mutate(marathon = case_when(
    marathon == "NYC" ~ "NY",
    marathon == "Grandmas" ~ "D",
    marathon == "Boston" ~ "B",
    marathon == "Twin Cities" ~ "TC"
  ))

#Change column names to match course_record_project column names
colnames(marathon_dates)[colnames(marathon_dates) == "marathon"] ="Race"
colnames(marathon_dates)[colnames(marathon_dates) == "year"] ="Year"

# Change formatting of the dates
marathon_dates <- marathon_dates %>%
  mutate(date = as.Date(date, format = "%Y-%m-%d"))

# Combine the marathon dates datframe to my current dataframe by using left_join
course_record_project1 <- course_record_project1%>%
  left_join(marathon_dates, by = c("Race", "Year"))

```

```

# Change the marathon variable to Race to match corresponding data and change race names
aqi_values <- aqi_values %>%
  rename(Race = marathon) %>%
  mutate(
    Race = case_when(
      Race == "NYC" ~ "NY",
      Race == "Grandmas" ~ "D",
      Race == "Boston" ~ "B",
      Race == "Twin Cities" ~ "TC"
    ),
    date = as.Date(date_local, format = "%Y-%m-%d"),
    Year = as.numeric(format(date, "%Y"))
  ) %>%
  select(-date_local) #Remove the date_local variable

# calculate average ozone ppm (8-hour avg)
avg_ppm <- aqi_values %>%
  filter(units_of_measure == "Parts per million",
        sample_duration == "8-HR RUN AVG BEGIN HOUR") %>%
  group_by(Race, Year, date) %>%
  summarize(avg_ppm = mean(arithmetic_mean, na.rm = T)) %>%
  ungroup()

# Merge data_frame to current dataframe
course_record_project1 <- course_record_project1 %>%
  left_join(avg_ppm, by = c("Race", "Year", "date"))

weather_parameters<-course_record_project1%>%
  select(Gender, `Age (yr)`, Runtimes, `Dry bulb Temp C`, `Wet bulb Temp C`,
`Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`, `Dew Point in C`,
`Wind Speed`, `Wet Bulb Globe Temp`, age_ranges, Flag)

# Flag Conditions on Gender
flag_exam <- ggplot(weather_parameters, aes(x = Gender, y = Runtimes, fill = Flag)) +
  geom_boxplot() + # Use geom_boxplot for creating a boxplot
  theme_bw() + # Black and white theme for a clean plot
  scale_fill_manual(values = c(
    "White" = "white", # Replace these with the Flag categories and the colors you want to assign
    "Green" = "green",
    "Black" = "black",
    "Yellow" = "yellow", # Add more colors for other Flag categories as needed
    "Red" = "red"
  )) +
  labs(
    title = "Marathon Performance by Flag Conditions Stratified by Gender",
    x = "Gender",
    y = "Marathon Performance (Runtimes)",
    fill = "Flag Color" # Title for the legend

```

```

) +
theme((legend.position = "right"),
      plot.title = element_text(hjust = 0.5, size = 12)) # Center and set title size to 12

flag_exam
# Flag conditions by Age Ranges
flag_age_ranges<- ggplot(weather_parameters, aes(x = age_ranges, y = Runtimes, fill = Flag)) +
  geom_boxplot() + # Use geom_boxplot for creating a boxplot
  theme_bw() +      # Black and white theme for a clean plot
  scale_fill_manual(values = c(
    "White" = "white",   # Replace these with the Flag categories and the colors you want to assign
    "Green" = "green",
    "Black" = "black",
    "Yellow" = "yellow", # Add more colors for other Flag categories as needed
    "Red" = "red"
  )) +
  labs(
    title = "Marathon Performance by Flag Conditions Stratified by Age Ranges",
    x = "Age Ranges",
    y = "Marathon Performance (Runtimes)",
    fill = "Flag Color" # Title for the legend
  ) +
  theme(
    (legend.position = "right"),
    plot.title = element_text(hjust = 0.5, size = 12))

flag_age_ranges

# Select only numeric columns from the data frame for correlation plot
airquality<- course_record_project1%>%
  select(Runtimes, `Dry bulb Temp C` , `Percent Relative Humidity` , `Black Globe Temp C` ,
         `Wind Speed` , `Dew Point in C` , `Solar Radiation` , `Age (yr)` , `Wet Bulb Globe Temp` ,
         avg_ppm)

#Compute the correlation matrix
cor_matrix2 <- cor(airquality, use = "complete.obs") # Use complete.obs to ignore NAs

```

```

# Melt the correlation matrix for ggplot2
library(reshape2)
cor_data2<- melt(cor_matrix2)

# Visualize the Correlation Plot of the the Weather Parameters
ggplot(data=cor_data2, aes(x= Var1, y= Var2, fill=value))+ 
  geom_tile(color= "white")+
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  scale_fill_gradient2(low= "hotpink", high="royalblue", mid = "white",
  midpoint=0,
  limit=c(-1,1),
  space="Lab",
  name="Correlation")+
  labs(title="Weather Parameters with the Largest Impact on Marathon Performance Correlation Plot")+
  theme_minimal(base_family="Times")+
  theme(axis.text.x= element_text(angle =45, vjust= 1, hjust = 1),
  plot.title = element_text(hjust = 0.5))

# Dry bulb plot
dry_bulb_plot<-ggplot(environmental_conditions, aes(x = `Dry bulb Temp C`, y = Runtimes, color = Gender))+
  geom_point(alpha=.1) + # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confidence interval
  facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
  theme_classic() +
  labs(title = "Marathon Performance vs. Dry Bulb Temperature by Age Ranges",
       x = "Dry Bulb Temperature (°C)",
       y = "Marathon Performance (Percent CR Seconds)")+
  scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))

dry_bulb_plot

# Humidity Percentages Plot
relative_percent_humidity<- ggplot(environmental_conditions, aes(x = `Percent Relative Humidity`,
  y = Runtimes, color = Gender)) +
  geom_point(alpha=.1) + # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confidence interval
  facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
  theme_classic() +
  labs(title = "Marathon Performance vs. Relative Humidity Stratified by Age Range",
       x = "Percent Relative Humidity %",
       y = "Marathon Performance (Runtimes)")+
  scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))

relative_percent_humidity

# Black Globe Temperature
black_globe_temp_graph<- ggplot(environmental_conditions, aes(x = `Black Globe Temp C`,
  y = Runtimes, color = Gender)) +
  geom_point(alpha=.1) + # Scatter plot

```

```

        geom_smooth(method = "lm", formula = y ~ x, color = "black") +
        # Linear regression line with confidence interval
        facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
        theme_classic() +
        labs(title = "Marathon Performance vs. Black Globe Temperature Stratified by Age Groups",
             x = "Black Globe Temp  (°C)",
             y = "Marathon Performance (Runtimes)")+
        scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))

black_globe_temp_graph

# Wet Bulb Temperature
wet_bulb_graph<-ggplot(environmental_conditions, aes(x = `Wet Bulb Globe Temp`,
            y = Runtimes, color = Gender)) +
    geom_point(alpha=.1) +
    geom_smooth(method = "lm", formula = y ~ x, color = "black") +
    # Linear regression line with confidence interval
    facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
    theme_classic() +
    labs(title = "Marathon Performance vs. Wet Bulb Globe Temp Stratified by Age Groups",
         x = "Wet Bulb Globe Temp (°C)",
         y = "Marathon Performance (Runtimes)")+
    scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))

wet_bulb_graph

#Solar Radiation Graph
solar_radiation_graph<-ggplot(environmental_conditions, aes(x = `Solar Radiation`,
            y = Runtimes, color = Gender)) +
    geom_point(alpha=.1) + # Scatter plot
    geom_smooth(method = "lm", formula = y ~ x, color = "black") +
    # Linear regression line with confidence interval
    facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
    theme_classic() +
    labs(title = "Marathon Performance vs. Solar Radiation Stratified by Age Groups",
         x = "Solar Radiation",
         y = "Marathon Performance (Runtimes)")+
    scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))

solar_radiation_graph

black_globe_temp_graph <- ggplot(environmental_conditions, aes(x = `Black Globe Temp C`, y = Runtimes,
            geom_point(alpha=.1) + # Scatter plot
            geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confidence interval
            facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
            theme_classic() +
            labs(title = "Marathon Performance vs. Black Globe Temperature Stratified by Age Groups",
                 x = "Black Globe Temp  (°C)",
                 y = "Marathon Performance (Runtimes)") +
            # Use scale_color_manual to set custom colors for Male and Female
            scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))

```

black_globe_temp_graph

table_summary