Impact of Environmental Conditions on Marathon 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 conditions 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`, `2` to `NY`, `3` to `TC`, and `4` to `D`. Additionally, the binary outcomes for `Gender` were modified to align across datasets, with `0` recoded as `F` for female and `1` as `M` for male.

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: Runtimes = Race_Seconds * (1 + (Percent CR / 100)).

Missingness

To begin this analysis, I got the sum of missing values from the dataset by columns. From this analysis, it was observed that certain weather paramters and environmental conditions were missing for particular marthons.Weather parameters like `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 specifically in Chicago 2011, Grandma's Duluth 2012, New York City 2018, and Twin Cities 2011 . 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 97.8% of the data remains well-represented.In the aqi_values dataset there was 3.9% missing in the data coming from the `Race` and `aqi` variables, but we have 96.1% present. Therefore,these missing data properties were remove from further analysis.

Exploratory Data Analysis

In this analysis, we will utilize line plots to examine trends in runtimes (performance) across various ages, highlighting the relationship between male and female marathon performances. Additionally, we will analyze summary statistics of runtimes across different age ranges to identify the age groups with the fastest (highest-performing) runtimes. Correlation plots will be used to assess the strength and direction of relationships among variables, offering a comprehensive understanding of their interdependencies.Futhermore,spline and generalized additive regression models will be employed to investigate the relationship between runtimes and other key variables, providing insights into potential predictors that impact marathon runner's performance.

Results

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

In the **Marathon Performance Summary Table by Age Ranges,** several key statistics highlight the performance trends of marathon runners across different age groups, including the number of participants, mean runtimes, median runtimes, and interquartile ranges (IQRs). The marathon runners were categorized into seven age groups, with a relatively balanced distribution of participants across these groups.

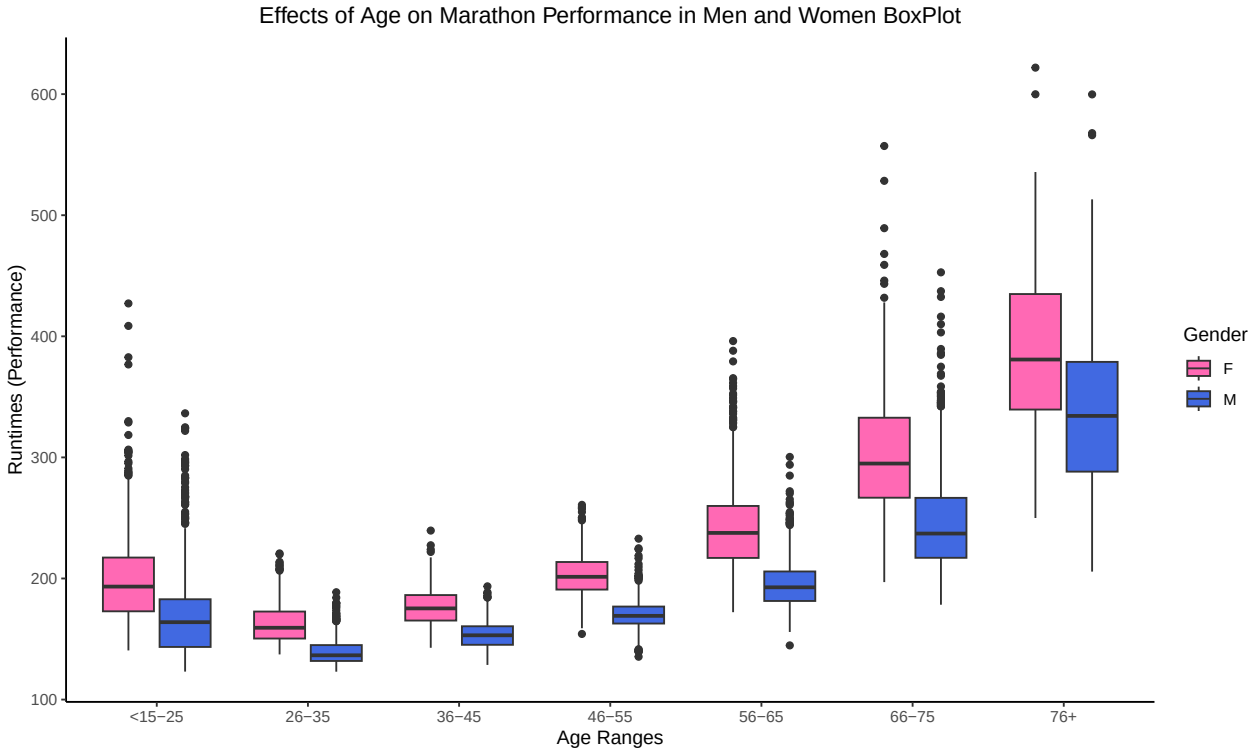
The **Effects of Age on Marathon Performance in Men and Women Boxplot** figure reveals that younger runners, specifically those aged 15–25, tend to achieve better runtimes. This is evidenced by lower median values and a tighter IQR. Despite their strong performance, this age group exhibits more variability, with a higher prevalence of outliers for both genders. The summary table indicates that this group has a mean runtime of 184 minutes and a median runtime of 177 minutes, with a relatively small IQR of 49. These results suggest that younger runners perform significantly better compared to older age groups, albeit with greater variability.

The 26–35 and 36–45 age groups demonstrate the best overall marathon performance. According to the boxplots, both men and women in these age ranges achieve lower runtimes with reduced variability, as indicated by shorter whiskers and fewer outliers. The summary table further confirms this trend, with median runtimes declining from 148 minutes for the 26–35 group to 163 minutes for the 36–45 group. The IQR remains relatively small (25) for both groups, reflecting consistent performance. Men consistently outperform women in these age ranges, as indicated by the lower median runtimes observed in the blue boxplots.

In contrast, the oldest age groups (66–75 and 76+) exhibit the longest runtimes and the greatest performance variability. The boxplots for these groups show much larger IQRs, particularly for women, highlighting significant variation in performance. The summary table indicates that mean and median runtimes increase notably for these groups: 269 and 259 minutes, respectively, for the 66–75 group, and 351 and 345 minutes for the 76+ group. The IQR expands substantially in the 76+ group, reaching 95, which underscores the broad range of abilities within this age range. The boxplot further demonstrates that men continue to outperform women even in older age groups. However, the performance gap between genders becomes increasingly pronounced after age 66, particularly in the 76+ group.

Marathon Performance Summary Table by Age Ranges

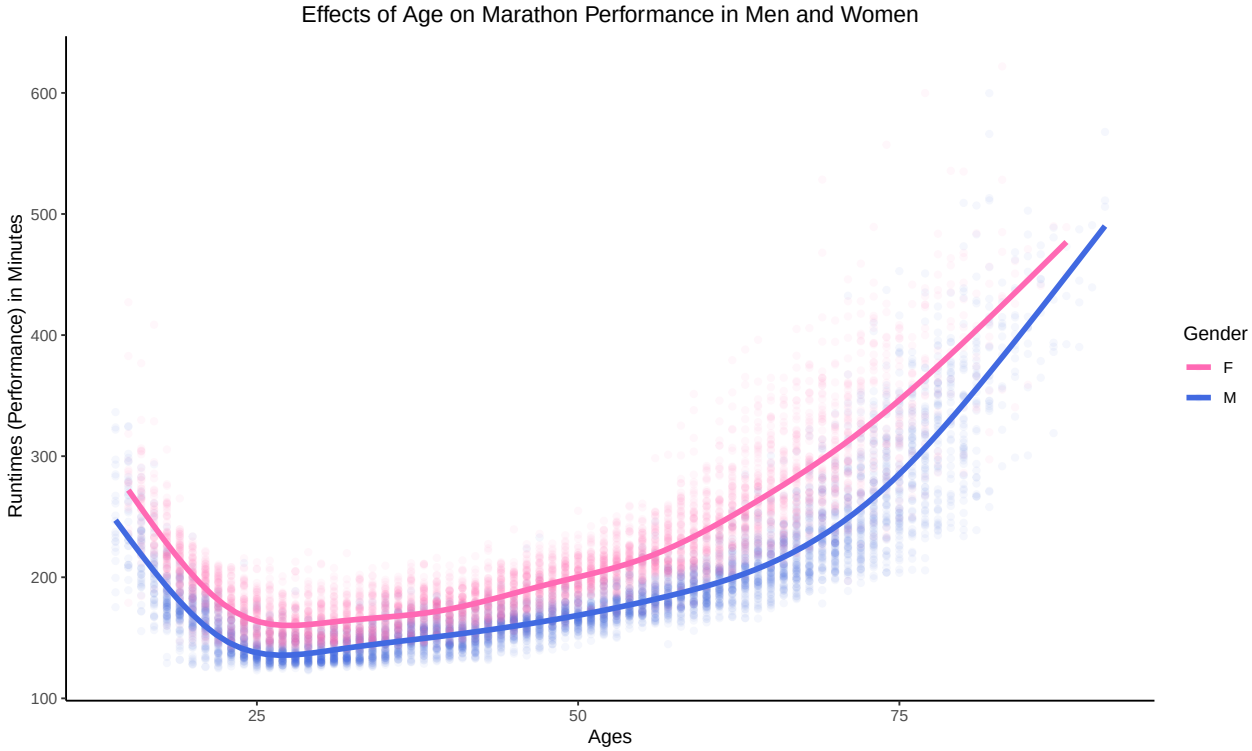
Age Ranges	Marathon Runners	Mean Runtimes	1Q	Median Runtimes	3Q	IQR
<15-25	1736	184	156	177	204	49
26-35	1840	151	137	148	161	25
36-45	1840	165	152	163	176	25
46-55	1840	186	169	183	202	33
56-65	1820	219	192	211	240	48
66-75	1463	269	228	259	299	71
76+	534	351	299	345	394	95



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 approximately around 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.

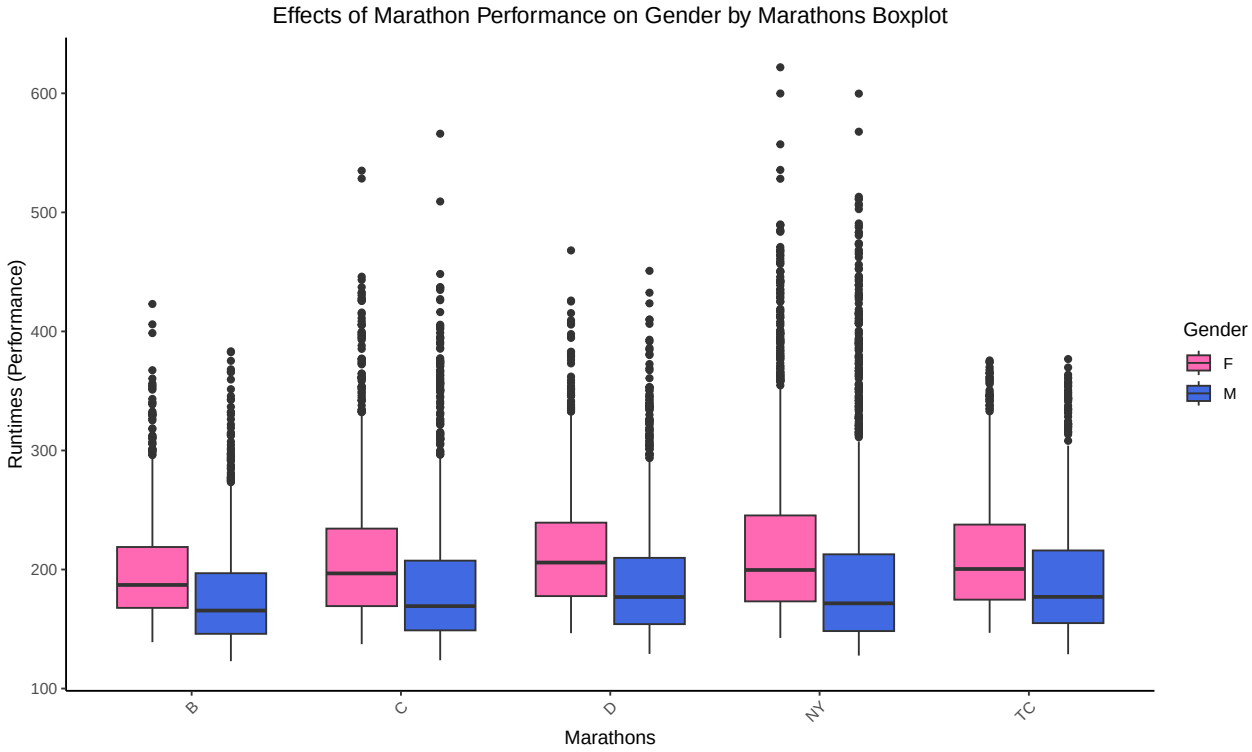


Based on the **Effects of Marathon Performance on Gender by Marathons Boxplot**, I examined the performances of men and women in each marathon. In the Boston Marathon, the men had the best overall performance, with a median runtime of 166 minutes (~2 hours and 26 minutes). The women's performance in this marathon was the best compared to their performance in other races, with a median runtime of 187 minutes (~3 hours and 7 minutes). The average runtime for men in this marathon was 178 minutes, while for women, it was 199 minutes.

Across all marathon races, men consistently had faster runtimes compared to women. The second-best performance by race and gender was observed in the Chicago Marathon. In this marathon, men had a median runtime of 169 minutes and an average runtime of 188 minutes, while women had a median runtime of 197 minutes and an average runtime of 212 minutes. The worst performance runtimes were observed during Duluth Grandma's Marathon, where the average runtime for women was 217 minutes and 192 minutes for men. The median runtime in this marathon was 206 minutes for women and 177 minutes for men.

Marathon Performance Summary by Gender and Marathon

Gender	Marathon Runners	Mean Runtimes	1Q	Median Runtimes	3Q	IQR
B						
F	984	199	168	187	219	51
M	1104	178	146	166	197	51
C						
F	1150	212	169	197	234	65
M	1277	188	149	169	207	59
D						
F	880	217	178	206	239	62
M	1004	192	154	177	210	56
NY						
F	1337	221	173	200	245	72
M	1462	198	148	172	213	64
TC						
F	867	212	175	200	238	63
M	1008	191	155	177	216	61

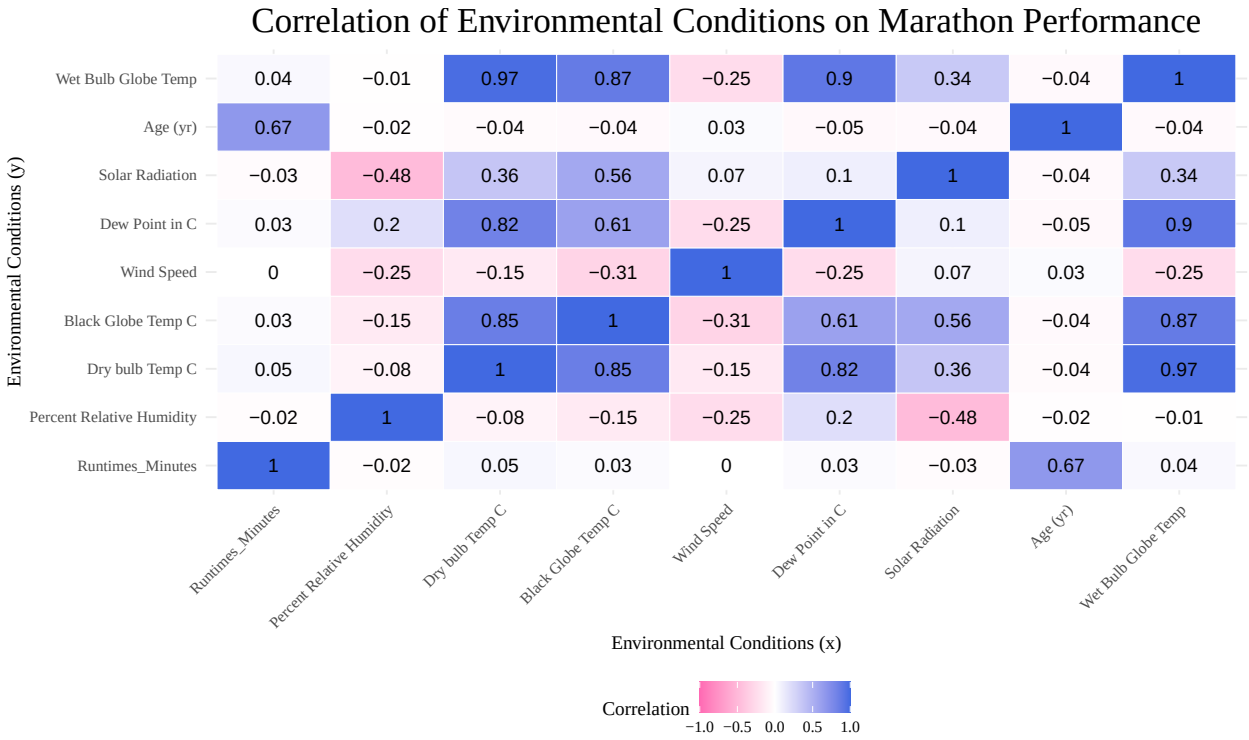


Overall Results to Aim 1: Examine effects of increasing age on marathon performance in men and women.

In Aim 1, we examined the effects of increasing age on marathon performance in men and women. Overall, men and women performed best in the age group of 26-35 years. As age increases beyond this range, there is a noticeable decline in performance, with the 76+ age group exhibiting significantly longer runtimes. This decline in performance with increasing age may be due to physiological factors such as reduced cardiovascular efficiency, decreased muscle strength, and slower recovery times, all of which are associated with aging.

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 spline model was conducted to demonstrate the relationship and impact of environmental conditions on marathon performance, considering differences across age and gender. This model was taken into consideration due to the non-linearity of the Runtimes to Ages in our model. Therefore, a spline regression was used with knots of `(20,40,60,80)`. The outcome variable is the marathon runners' runtimes in minutes, analyzed in conjunction with environmental conditions, gender, and age. Upon examining the model, I conducted a correlation plot to see the relationship of the weather parameters amongst each other. In this process, I noticed that `Wet Bulb Globe Temp` has multiple high correlations with other weather parameters like `Dry Bulb Temp C`(.97), `Black Globe Temp C` (.87), and `Dew Point in C`(.9). Due to this correlation, the three associates of `Wet Bulb Globe Temp` were removed to reduce multicollinearity in my model.



Following the model explanation, I found that the highlighted blue parameters are significant parameters. `Percent Relative Humidity` (p=0.00000) and `Wet Bulb Globe Temp C` (p=0.00805) demonstrated significant weather parameters that impacts marathon performance. The interaction of `GenderM` with "Percent Relative Humidity" (p = 0.00004) highlights that males are disproportionately affected by humidity.The results also indicate that race parameters such as `RaceC` (p= 0.00000), `RaceD` (p= 0.00001), `RaceNY` (p =0.00000), and `RaceTC` (p = 0.00000) are statistically significant predictors of performance. Therefore, maybe the plains or cities which deals with environmental factors may contribute to marathon runners' performance.

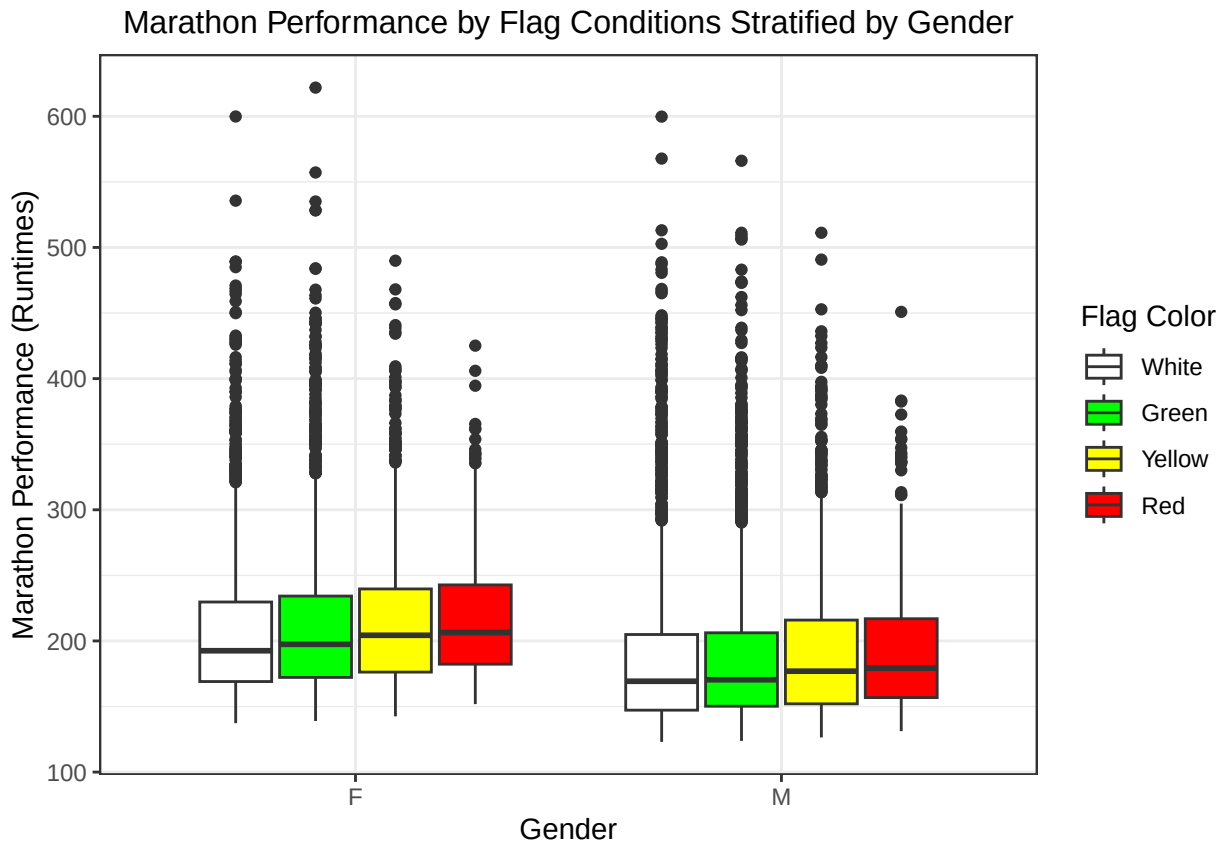
Other significant weather parameters analysis was conducted in the Supplementary Materials below.

Table 1: Spline Model:Impact of Environmental Conditions on Marathon Performance by Age and Gender

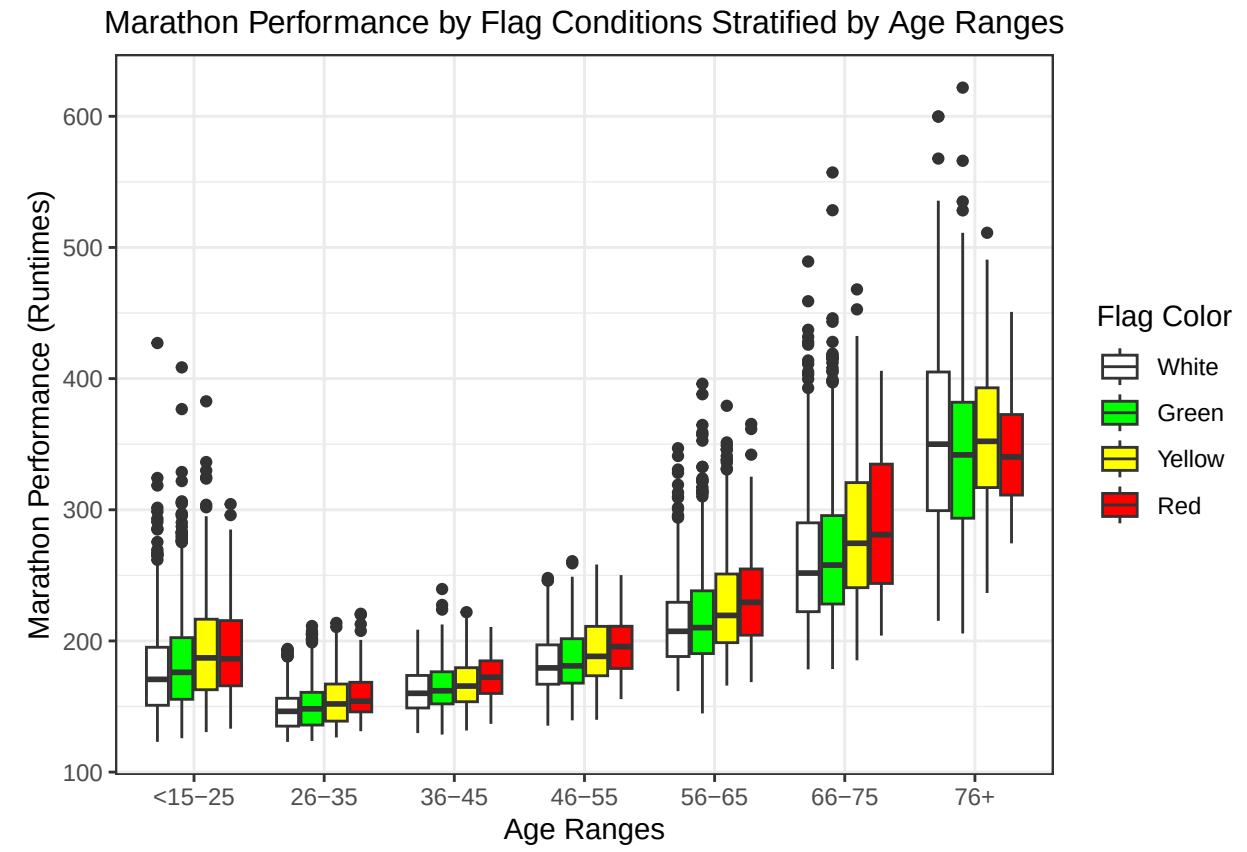
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**264.45606**	**3.85439**	**68.61172**	**0.00000**
ns('Age (yr)', knots = knots)1	**-93.55248**	**2.23346**	**-41.88679**	**0.00000**
ns('Age (yr)', knots = knots)2	**-64.00214**	**3.39704**	**-18.84058**	**0.00000**
ns('Age (yr)', knots = knots)3	**121.84456**	**3.10350**	**39.26036**	**0.00000**
ns('Age (yr)', knots = knots)4	**72.06124**	**8.29827**	**8.68389**	**0.00000**
ns('Age (yr)', knots = knots)5	**312.39870**	**8.54944**	**36.54027**	**0.00000**
GenderM	**-43.61873**	**5.24134**	**-8.32206**	**0.00000**
RaceC	**13.58360**	**1.19121**	**11.40323**	**0.00000**
RaceD	**19.34575**	**1.24373**	**15.55456**	**0.00000**
RaceNY	**7.09351**	**1.40537**	**5.04744**	**0.00000**
RaceTC	**16.92192**	**1.33471**	**12.67838**	**0.00000**
'Percent Relative Humidity'	**-0.11820**	**0.01627**	**-7.26676**	**0.00000**
'Wind Speed'	0.16240	0.09770	1.66223	0.09650
'Solar Radiation'	-0.00636	0.00328	-1.93994	0.05241
'Wet Bulb Globe Temp'	**0.82767**	**0.31228**	**2.65037**	**0.00805**
ns('Age (yr)', knots = knots)1:GenderM	**23.76607**	**3.04286**	**7.81044**	**0.00000**
ns('Age (yr)', knots = knots)2:GenderM	-4.06098	4.47123	-0.90825	0.36377
ns('Age (yr)', knots = knots)3:GenderM	**-49.06671**	**3.89516**	**-12.59684**	**0.00000**
ns('Age (yr)', knots = knots)4:GenderM	**30.40886**	**10.03945**	**3.02894**	**0.00246**
ns('Age (yr)', knots = knots)5:GenderM	**34.02852**	**9.64127**	**3.52946**	**0.00042**
GenderM:RaceC	-2.19536	1.64177	-1.33719	0.18119
GenderM:RaceD	**-4.04593**	**1.70816**	**-2.36860**	**0.01787**
GenderM:RaceNY	0.29013	1.93566	0.14989	0.88086
GenderM:RaceTC	-1.75479	1.83199	-0.95786	0.33815
GenderF:'Dry bulb Temp C'	-0.21438	0.29221	-0.73365	0.46318
GenderM:'Dry bulb Temp C'	0.15864	0.27347	0.58008	0.56187
GenderM:'Percent Relative Humidity'	**0.09170**	**0.02236**	**4.10167**	**0.00004**
GenderM:'Wind Speed'	0.00549	0.13454	0.04082	0.96744
GenderM:'Solar Radiation'	0.00825	0.00451	1.82881	0.06746
GenderM:'Wet Bulb Globe Temp'	-0.58021	0.42797	-1.35574	0.17521

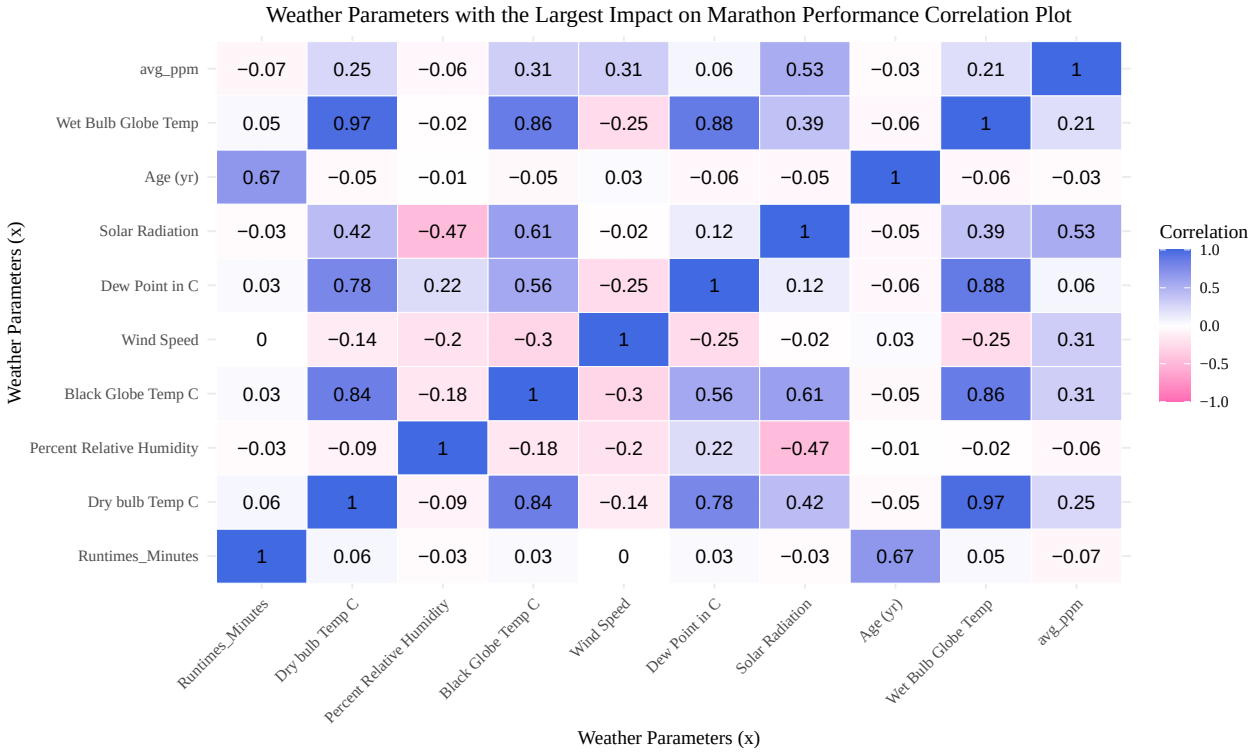
Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

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.



According **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.





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).

Using a Generalized Additive Model (GAM), I observed an extension of linear models that allows for flexible modeling of non-linear relationships between predictors and a response variable.In thus analysis, we observe the weather and flag conditions that has the most impact on marathon runners' performance. The flag conditions that has largest impact on marathon performers are Red (p=5.04e-06) and Yellow Flag (p=2.24e-08) which indicates that harsher environmental conditions increase runtimes.

`Avg_ppm` shows a strong and significant negative association with runtime (p =5.30e-14), indicating that increased `avg_ppm` substantially reduced runtimes, likely reflecting improved environmental conditions. This indications means that improved air quality (higher avg_ppm values) significantly reduces runtimes, potentially due to better oxygen availability or reduced environmental strain on runners Ase environmental conditions, air quality, and gender differences on marathon performance.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Runtimes_Minutes ~ Flag + avg_ppm + Gender
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  222.593      2.111 105.428  < 2e-16 ***
## FlagGreen      2.304      1.504   1.532    0.126
## FlagYellow     9.828      1.756   5.598 2.24e-08 ***
## FlagRed       13.590      2.976   4.566 5.04e-06 ***
## avg_ppm     -486.126     64.500  -7.537 5.30e-14 ***
## GenderM      -22.625      1.282 -17.648  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.0434   Deviance explained = 4.39%
## GCV = 3542.6  Scale est. = 3540.2    n = 8646
```

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

From the boxplot illustrating **Marathon Performance Stratified by Gender and Flag Conditions**, women (F) appear to be more impacted by adverse weather conditions compared to men (M). Under green (moderate), yellow (caution), and red (high-risk) flag conditions, women's runtimes exhibit a wider distribution, with a higher median and greater variability compared to their performance under white (favorable) conditions. This suggests that adverse weather conditions exacerbate challenges for women more significantly, potentially due to physiological differences in thermoregulation or endurance in extreme environments.

Men, on the other hand, show relatively less variation across flag conditions, with smaller differences in median and interquartile ranges when transitioning from favorable (white) to adverse (red) conditions. While men are still affected by harsh weather, their performance tends to be more consistent across different flag conditions.

From the boxplot, it is evident that the **Impact of Flag Weather Parameters on Marathon Performance** varies across age ranges. The flag colors—indicating different weather conditions (e.g., white for favorable, green for moderate, yellow for caution, and red for high risk)—appear to have the most pronounced effect on older age groups. Specifically, runners aged **66–75** and **76+** show the widest variation in runtimes across flag conditions, with slower runtimes under yellow and red flags compared to white and green. This suggests that older runners are more adversely affected by less favorable weather conditions, likely due to reduced physiological resilience and greater susceptibility to heat stress or other weather-related factors.

In contrast, younger runners, particularly those aged **26–35**, exhibit relatively smaller performance differences across flag conditions. This age group consistently demonstrates the fastest runtimes, with minimal deviation between favorable (white/green) and unfavorable (yellow/red) weather conditions. Runners in this age group seem better able to maintain their performance despite adverse weather, possibly due to better thermoregulatory capabilities and greater overall fitness. However, as age increases, the influence of adverse weather conditions becomes progressively more significant, underscoring the need for targeted safety measures for older marathon participants during challenging weather conditions.

Among the flag conditions, Red (p = 5.04e-06) and Yellow (p = 2.24e-08) flags have the most substantial influence, indicating that harsher conditions, such as elevated temperature and humidity, lead to increased runtimes. Additionally, avg_ppm shows a strong and significant negative association with runtimes (p = 5.30e-14), suggesting that improved air quality enhances performance by reducing runtimes. This improvement is likely due to better oxygen availability and reduced physiological strain on runners under cleaner air conditions. Together, these findings emphasize the critical roles of environmental conditions, air quality, and gender in shaping marathon performance outcomes.

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 highlights the significant impact of environmental conditions on marathon performance across genders and age groups. According to *Aerobic Performance is Degraded Despite Modest Hypothermia in Hot Environments*, "Endurance exercise performance is degraded with increasing environmental temperature, and the decline in performance associated with warmer temperatures is magnified in longer-distance events such as the marathon foot race" (Ely, B. R., et al.). Research indicates that marathon runners prefer cooler, less humid conditions because the body can more effectively regulate heat in these environments. Running in cooler weather reduces the risk of heat-related illnesses, such as heat exhaustion or heat stroke, which are more prevalent in warmer conditions. Additionally, cooler temperatures lessen the strain on the cardiovascular system, enabling runners to maintain their pace without a significant increase in heart rate, which is often observed in hotter conditions.

In my analysis, environmental factors such as Percent Relative Humidity and Wet Bulb Globe Temperature (WBGT) play a substantial role in marathon performance. Yellow and Red flag conditions, which signal more challenging weather, have the most significant impact on runners by increasing runtimes, while better air

quality is associated with improved performance. Runners typically perform best under Green flag conditions, which indicate safer weather with lower WBGT levels. This is reflected in the lower median runtimes observed under Green flags compared to more severe conditions marked by Red, White, and Yellow flags. As weather conditions transition to Red and Yellow flags, indicative of higher WBGT levels, a notable decline in performance is observed, with increased median runtimes. These harsher conditions, characterized by elevated temperatures and humidity, contribute to the performance drop. Moreover, the increased variability in performance under Red and Yellow flags suggests differing levels of heat tolerance among runners.

While both male and female runners exhibit similar trends in response to changing flag conditions, female runners consistently show slightly higher median runtimes under all conditions, indicating relatively slower performance. This highlights subtle yet consistent gender differences in marathon performance under varying environmental conditions.

Limtations

The limitations I want to highlight in this study include the non-linear relationship between age and runtimes. To address this limitation, I implemented a spline regression model, as it is better suited for handling non-linear relationships. Additionally, there was a significant indication of high variance inflation factors (VIF) among the environmental conditions, which led to issues of high correlations and multicollinearity. To resolve this problem, environmental conditions such as `Black Globe Temp C`, `Dry Bulb Temp C`, and `Dew Point in C` were removed to reduce the high correlations.

The Wet Bulb Globe Temperature (WBGT) was retained as it provides a more comprehensive measure by incorporating all the environmental conditions that were removed individually. This adjustment allowed for a more robust model that effectively accounted for environmental influences while reducing the impact of multicollinearity.

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.

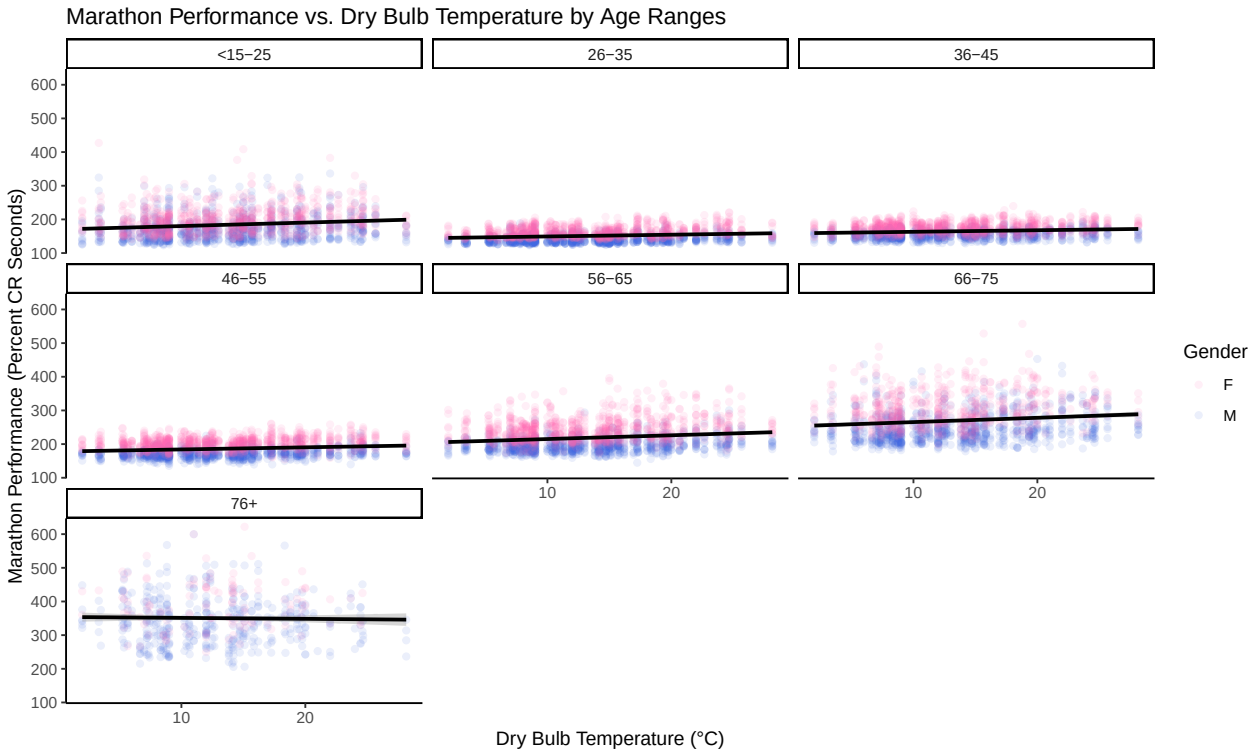
Marathon Performance vs. Solar Radiation Stratified by Age Groups

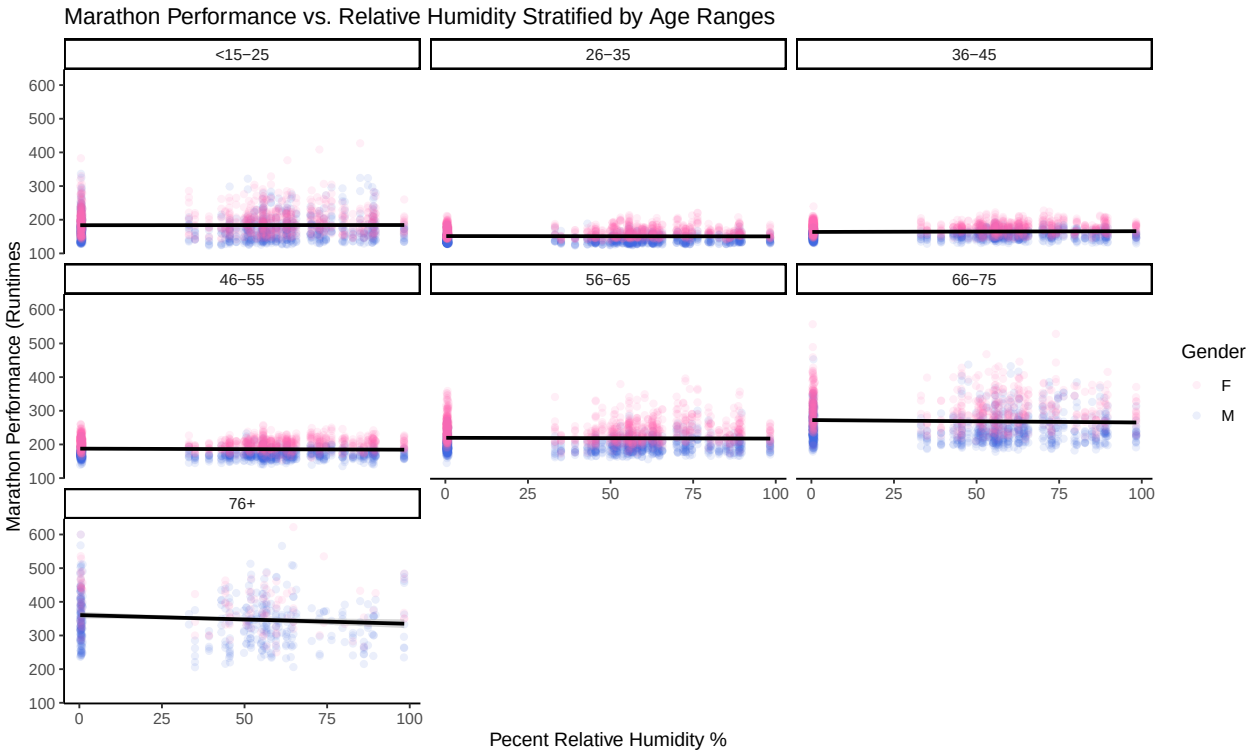
Across all age groups, there is little to no effect of solar radiation on marathon performance. The trend lines across different levels of solar radiation (ranging from 0 to 750 units) are mostly flat, indicating that

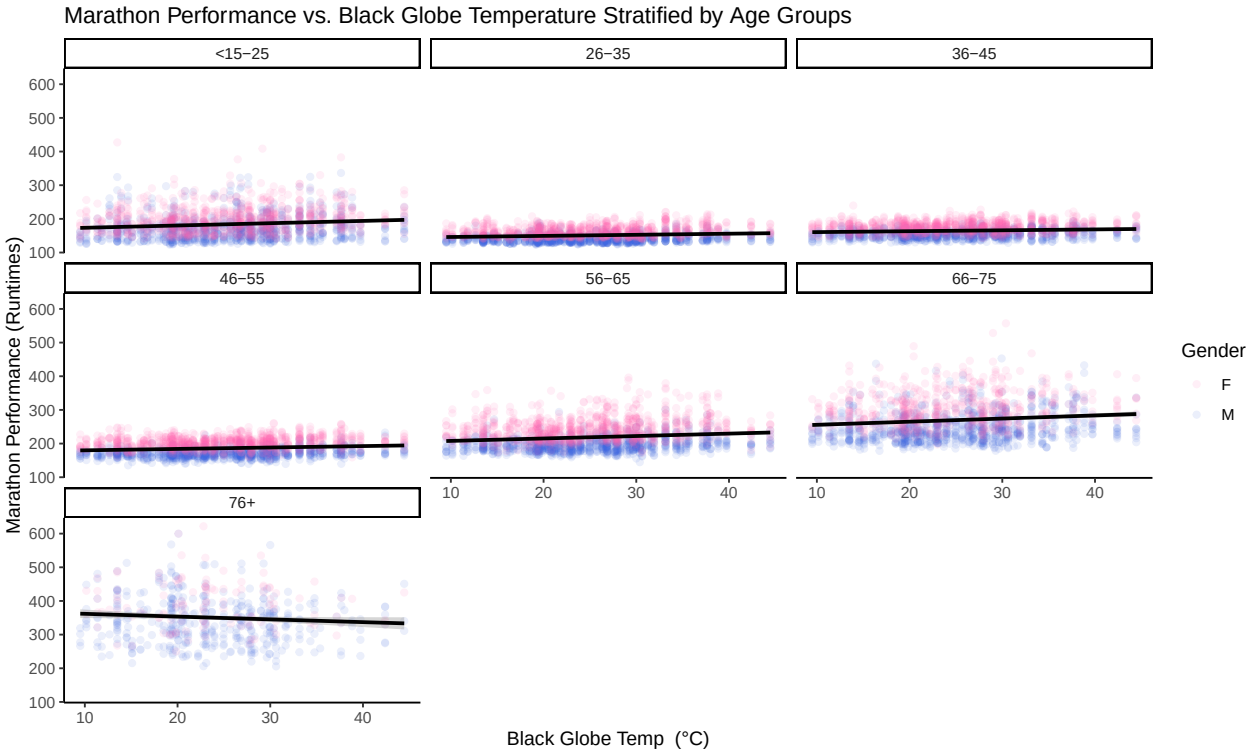
increased solar radiation does not have a strong impact on runtimes for either gender. This indicates that solar exposure alone does not significantly affect marathon performance.

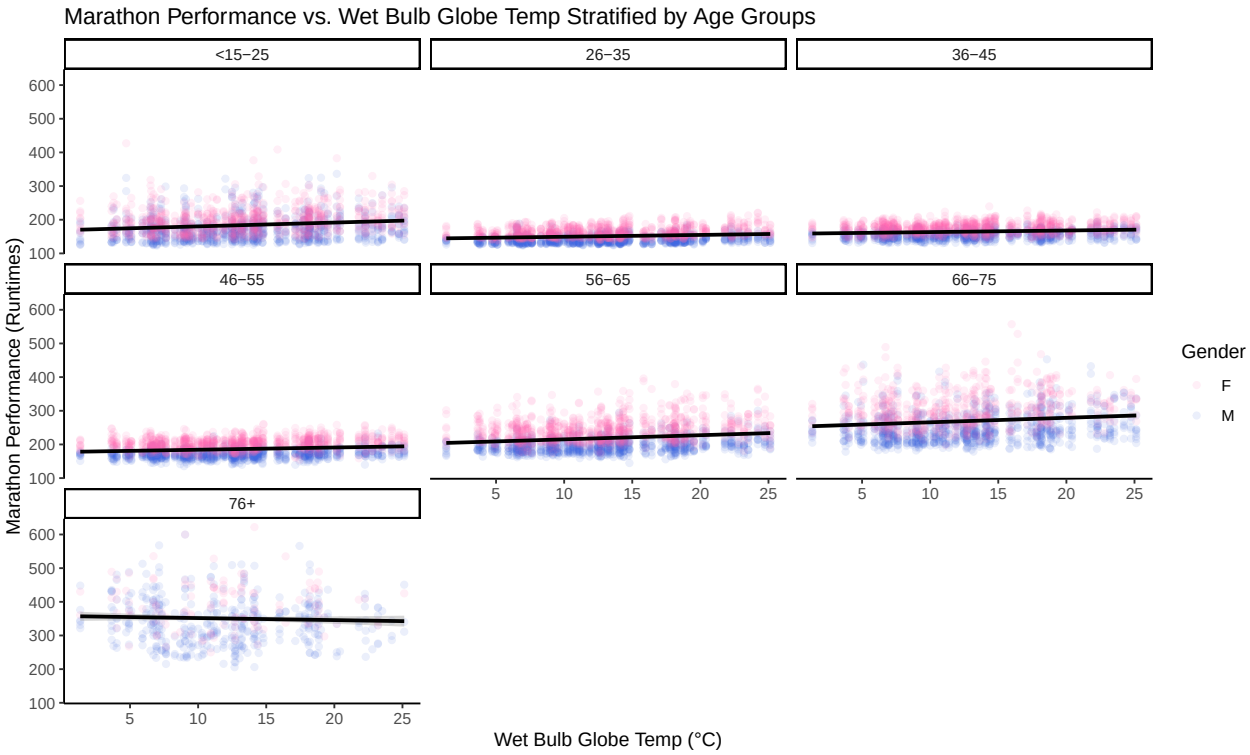
Marathon Performance vs. Black Globe Temperature Stratified by Age Groups

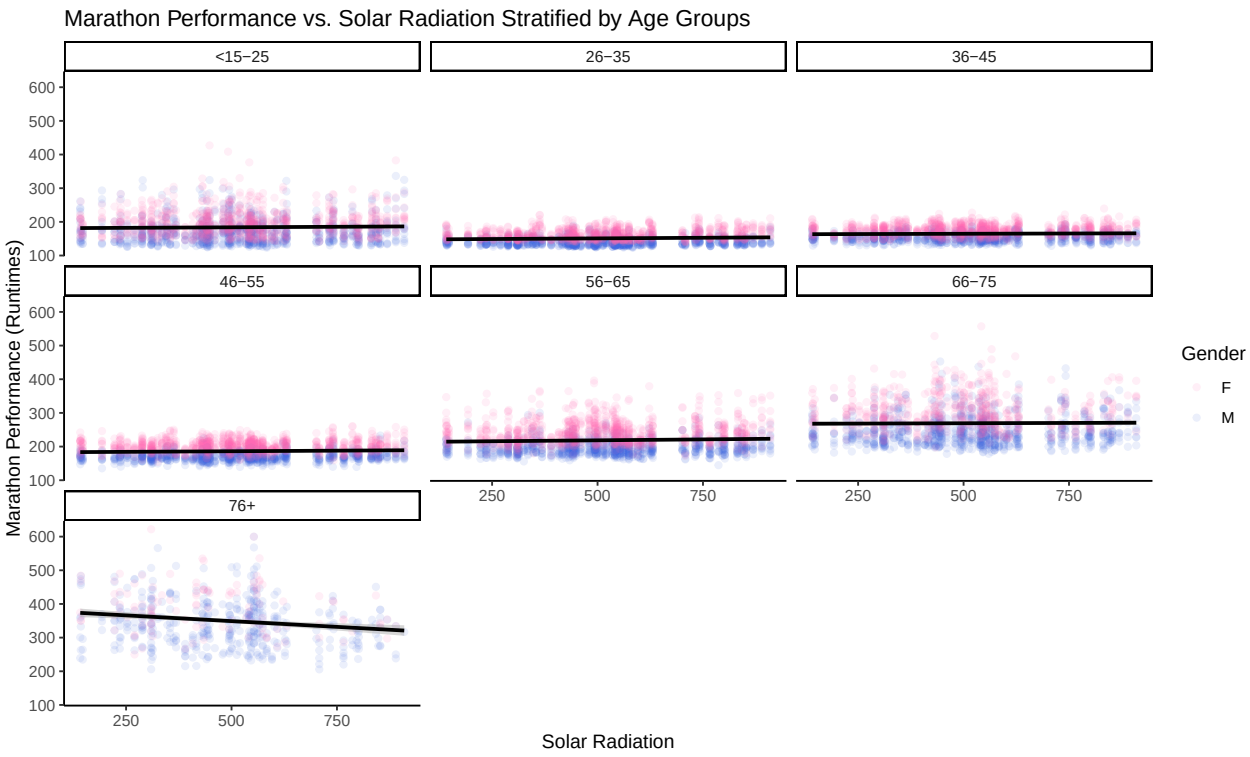
The slope of the trend lines across all age groups is generally flat, suggesting that Black Globe Temperature (BGT) does not have a strong impact on marathon performance, especially in younger age groups (<15-25, 26-35, 36-45). This indicates that younger runners are more resilient to changes in environmental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




Marathon Performance vs. Relative Humidity Stratified by Age Ranges
<15−25
26−35
36−45
46−55
56−65
66−75
76+
Marathon Performance (Runtimes
Pecent Relative Humidity %
Gender
F
M


Marathon Performance vs. Black Globe Temperature Stratified by Age Groups
<15−25
26−35
36−45
46−55
56−65
66−75
76+
Marathon Performance (Runtimes)
Black Globe Temp (°C)
Gender
F
M


Marathon Performance vs. Wet Bulb Globe Temp Stratified by Age Groups
<15−25
26−35
36−45
46−55
56−65
66−75
76+
Marathon Performance (Runtimes)
Wet Bulb Globe Temp (°C)
Gender
F
M


Marathon Performance vs. Solar Radiation Stratified by Age Groups
<15−25
26−35
36−45
46−55
56−65
66−75
76+
Marathon Performance (Runtimes)
Solar Radiation
Gender
F
M

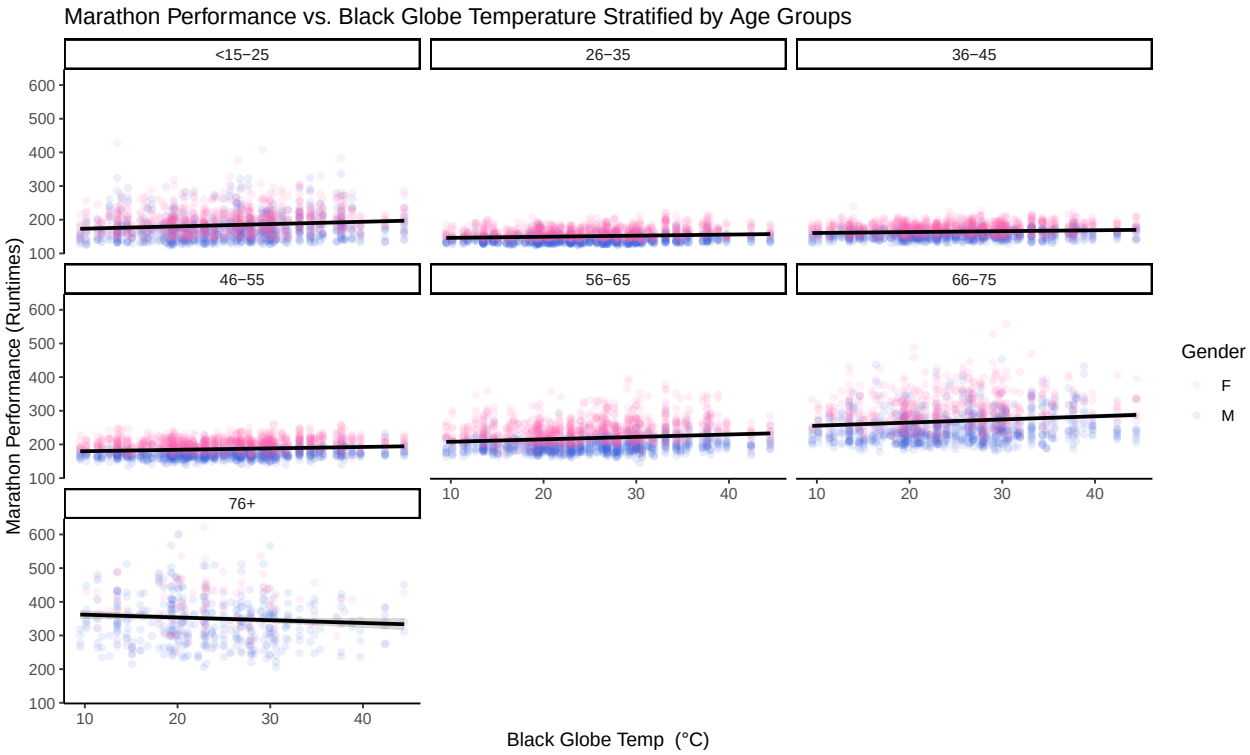

Marathon Performance vs. Black Globe Temperature Stratified by Age Groups
<15−25
26−35
36−45
46−55
56−65
66−75
76+
Marathon Performance (Runtimes)
Black Globe Temp (°C)
Gender
F
M

Table 2: 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.
Runtimes_Minutes	0	NA	NA
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 x Tw + 0.2 x Tg+ 0.1 xTd.
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)
library(car)
library(carData)
library(mgcv)



# 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()


#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 el
```

```r
# 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 minutes
course_record_project1 <- course_record_project1 %>%
  mutate(Runtimes = Race_Seconds * (1 + (`Percent CR` / 100)),
         Runtimes_Minutes= Runtimes / 60)




#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()

vis_miss(aqi_values)


#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) %>%
  dplyr::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","Numeric", "HMS/Numeric", "Numeric","Numeric"),
  Description= c("Race represents the marathons the participants competed, including the B=Boston Marath
                 C= Chicago Marathon, NY= New York City Marathon,T= Twin Cities Marathon (Minneapolis,M
                 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 humidi
                 moisture.",
                 "Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humi
                 "Percent Relative Humidity how much moisture is in the air compared to the maximum amou
                 "Black Globe Temp Celcius  indicates how hot it feels in direct sunlight.It considers t
                 "Solar Radiation in Watts per meter squared is the energy emitted by the sun, which tra
                 "Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pre
                 "Wind Speed in Km/hr.",
                 "Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed
                 "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
minimimum_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%>%
  dplyr::select(Race,Year, Gender, age_ranges, Runtimes_Minutes, `Age (yr)`)%>%
  group_by(Race, Gender, age_ranges)
```

```r
#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 wit
  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 Ag
  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
  facet_wrap(~ age_ranges) +
  theme_minimal()


#Create summary statistics table
marathon_performance_summary_table <-marathon_performance_by_age%>%
  group_by(age_ranges) %>%
```

```
  summarise(
    Count=n(),
    Mean= mean(Runtimes_Minutes),
    Q1_Runtime = quantile(Runtimes_Minutes, 0.25, na.rm = TRUE),
    Median_Runtime = median(Runtimes_Minutes, na.rm = TRUE),       # Median of Runtimes
     # Lower quartile (25th percentile)
    Q3_Runtime = quantile(Runtimes_Minutes, 0.75, na.rm = TRUE),    # Upper quartile (75th percentile)
    IQR_Runtime = IQR(Runtimes_Minutes, na.rm = TRUE) # IQR of Runtimes
  )

marathon_performance_summary_table %>%
  gt() %>%
  tab_header(
    title = "Marathon Performance Summary Table by Age Ranges"
  ) %>%
  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_Minutes, 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


# Line plot of the Marathon Runners Performance by Age
age_plot<-ggplot(marathon_performance_by_age, aes(x = `Age (yr)`, y = Runtimes_Minutes, 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) in Minutes") +
  scale_color_manual(values = c("F" = "hotpink", "M" = "royalblue")) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5) )



age_plot




# Create boxplot stratified by diffferent Races
ggplot(marathon_performance_by_age, aes(x = Race, y = Runtimes_Minutes, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Effects of Marathon Performance on Gender by Marathons Boxplot",
       x = "Marathons",
       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) )

marathon_performance_summary_table2 <-marathon_performance_by_age%>%
  group_by(Race, Gender) %>%
  summarise(
    Count=n(),
    Mean= mean(Runtimes_Minutes),
    Q1_Runtime = quantile(Runtimes_Minutes, 0.25, na.rm = TRUE),
    Median_Runtime = median(Runtimes_Minutes, na.rm = TRUE),       # Median of Runtimes
     # Lower quartile (25th percentile)
    Q3_Runtime = quantile(Runtimes_Minutes, 0.75, na.rm = TRUE),    # Upper quartile (75th percentile)
    IQR_Runtime = IQR(Runtimes_Minutes, na.rm = TRUE) # IQR of Runtimes
  )

marathon_performance_summary_table2 %>%
  gt() %>%
  tab_header(
    title = "Marathon Performance Summary by Gender and Marathon"
  ) %>%
  cols_label(
    Race = "Age Ranges",
    Race = "Race by Gender",
    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 Correlation plot
# Observing  and Including all the numeric variables
numeric_data <- course_record_project1%>%
  dplyr::select(Runtimes_Minutes, `Percent Relative Humidity`, `Dry bulb Temp C`, `Black Globe Temp C`,

# 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",
       x= "Environmental Conditions (x)",
       y= "Environmental Conditions (y)") +
  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

 # Create Dataframe to allocate for other environmental conditions
 environmental_conditions <- course_record_project1 %>%
  dplyr::select(Race, Gender, `Age (yr)`, age_ranges, Runtimes_Minutes, `Dry bulb Temp C`, `Wet bulb Ter
  `Wind Speed`, `Wet Bulb Globe Temp`)


library(splines)

# Define the knots
knots <- c(20, 40, 60, 80)  # Update to match your visualization purpose
```

```r
# Fit a spline model with appropriate knots

spline_model <- lm(Runtimes_Minutes ~
                     ns(`Age (yr)`, knots = knots) * Gender + # Interaction between spline(age) and Gen
                      Race +
                     `Percent Relative Humidity` +
                     `Wind Speed` +
                     `Solar Radiation` +
                     `Wet Bulb Globe Temp` +
                     `Race`: Gender+
                     `Dry bulb Temp C`:Gender +
                     `Percent Relative Humidity`:Gender +
                     `Wind Speed`:Gender +
                     `Solar Radiation`:Gender +
                     `Wet Bulb Globe Temp`:Gender,
                   data = environmental_conditions)



# Get the beta coeffiecients of the spline model
spline_table <- round(summary(spline_model)$coefficients, 5)


# Use kable to create the summary table
spline_table %>%
  kbl(caption = "Spline Model:Impact of Environmental Conditions on Marathon Performance by Age and Gen
      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(5, 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(15, bold = TRUE, color="black", background = "#87CEEB")%>%
  row_spec(16, bold = TRUE, color="black", background = "#87CEEB")%>%
  row_spec(18, bold = TRUE, color="black", background = "#87CEEB")%>%
   row_spec(19, bold = TRUE, color="black", background = "#87CEEB")%>%
  row_spec(20, bold = TRUE, color="black", background = "#87CEEB")%>%
   row_spec(22, bold = TRUE, color="black", background = "#87CEEB")%>%
   row_spec(27, bold = TRUE, color="black", background = "#87CEEB")



#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"))
      ) %>%
   dplyr::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"))
```

```r
weather_parameters<-course_record_project1%>%
  dplyr::select(Gender, Race, `Age (yr)`, Runtimes_Minutes, `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, `avg_ppm`)
# Ensure the Flag variable is a factor and order it by severity
weather_parameters$Flag <- factor(
  weather_parameters$Flag,
  levels = c("White", "Green", "Yellow", "Red", "Black")  # Order by severity
)


# Flag Conditions on Gender
flag_exam <- ggplot(weather_parameters, aes(x = Gender, y = Runtimes_Minutes, 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_Minutes, 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%>%
  dplyr::select(Runtimes_Minutes, `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",
       x="Weather Parameters (x)",
        y="Weather Parameters (x)"
       )+
  theme_minimal(base_family="Times")+
  theme(axis.text.x= element_text(angle =45, vjust= 1, hjust = 1),
    plot.title = element_text(hjust = 0.5))
```

```r
gam_model <- gam(Runtimes_Minutes ~ Flag+ avg_ppm+ Gender, data =weather_parameters)
summary(gam_model)
```

```r
# Dry bulb plot
 dry_bulb_plot<-ggplot(environmental_conditions, aes(x = `Dry bulb Temp C`, y = Runtimes_Minutes, color
  geom_point(alpha=.1) +  # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") +  # Linear regression line with confide
  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_Minutes, 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
                                    x = "Pecent 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_Minutes, 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
                           x = "Black Globe Temp  (°C)",
                           y = "Marathon Performance (Runtimes)")+
                           scale_color_manual(values = c("M" = "royalblue", "F" = "hotpink"))
```

```r
black_globe_temp_graph


 # Wet Bulb Temperature
 wet_bulb_graph<-ggplot(environmental_conditions, aes(x = `Wet Bulb Globe Temp`,
                       y = Runtimes_Minutes, 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_Minutes, 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_M
  geom_point(alpha=.1) +  # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") +  # Linear regression line with confide
  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
```