Effects of Age, Gender, and Environmental Factors on Marathon Performance

Morgan Cunningham

2024-10-01

Introduction

This project is a collaborative effort with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. Marathon performance is widely recognized to decline as environmental temperatures increase, particularly in endurance events like marathons, where thermoregulatory challenges play a critical role. Older adults are especially affected, as their ability to dissipate heat diminishes with age, making them more susceptible to the effects of heat stress (Knechtle et al., 2021). Additionally, gender-based differences in endurance performance and physiological responses to environmental conditions have been well-documented. For instance, men are more likely than women to slow their pace over the course of a marathon (Deanor et al., 2015), whereas older runners, regardless of gender, tend to maintain a more consistent pace compared to younger runners (Nikolaidis, 2019).

The impacts of environmental factors, such as temperature, humidity, solar radiation, and wind speed, are further complicated by age and gender. Research has shown that higher temperatures and humidity negatively influence the performance of masters athletes, as observed in the New York City Marathon (Knechtle et al., 2021). These findings highlight the need for a nuanced understanding of how individual and environmental factors interact to shape marathon outcomes.

The primary objective of this project is to assess the impact of environmental conditions on marathon performance across a broad spectrum of age and gender groups. Using a large dataset from several major marathons, spanning performances by individuals aged 14 to 85, the study incorporates detailed environmental data for each event.

The study will address the following aims:

AIM 1: Investigate the effects of increasing age on marathon performance in both men and women, including pace variance and performance trends.

AIM 2: Explore how various environmental factors influence marathon performance and examine whether these impacts vary across different age groups and genders.

AIM 3: Identify which weather parameters have the largest effect on marathon performance, hypothesizing that increased environmental temperatures and adverse weather conditions negatively impact overall performance.

Project Dataset

The dataset for this project includes marathon race results for men and women and detailed weather data from five major marathon events: Boston, Chicago, New York, Twin Cities (Minneapolis, MN), and Grandma's Marathon (Duluth, MN). These events span from 1993 to 2016, covering 17 to 24 years of performances. Weather data associated with each race includes key environmental parameters such as wet bulb globe temperature (WBGT), dry bulb temperature, humidity, solar radiation, and wind speed. Additionally,

the dataset captures individual performance metrics, such as finish times and their percentage difference from the course record (%CR), enabling a detailed analysis of top-performing runners across different age groups and genders. Preliminary analyses involved extracting the fastest finishing times among men and women for each year of age and comparing these to the course record.

To prepare for analysis, three data sets including the course records, the aforementioned marathon performance data, and AQI data were merged. The AQI data comes from the U.S. Environmental Protection Agency's Air Quality System, which provides detailed air quality information for various pollutants across the United States. We utilized the `{RAQSAPI}` package to interact with the AQS API, enabling programmatic retrieval of air quality indices, pollutant concentrations, and related environmental parameters directly from the EPA's database. Specifically, we collected AQI data for marathon events held in the five cities of interest: Boston, Chicago, New York City, Minneapolis, and Duluth. The resulting dataset includes information on key air quality parameters such as PM2.5, PM10, and ozone, corresponding to the years and locations within our study. To actually merge our AQI data,we calculated the average AQI in parts per million for each race, year, and date. We then joined the summarized AQI data with our final marathon data set. To merge the course records and marathon performance data, variable names were standardized for clarity, and race times were converted to seconds to facilitate accurate comparisons. Additionally, an entirely new data set was written with these changes, ensuring the integrity and usability of the data throughout the analysis.

Upon review, we observe that there are missing values across several weather-related variables within four major races. We see that missing values are not random but tied to specific races and years, suggesting that for certain races in specific years environmental data may not have been recorded or reported properly due to equipment malfunction or inconsistent data collection practices during those events. More specifically, Chicago 2011, Grandma's 2012, New York City 2018, and Twin Cities 2011 are all missing Dry Bulb Temperature, Wet Bulb Temperature, Humidity, Black Globe Temperature, Solar Radiation, Dew Point, Wind, and Wet Bulb Globe Temperature for each observation. In total, approximately 4.25% of the total data set is missing. Since the overall percentage of missing values is below the commonly accepted threshold of 5%, it will have a negligible impact on the results. Therefore, the decision to proceed with observed data only was made.

AIM 1

Investigate the effects of increasing age on marathon performance in both men and women.

We start be creating a summary table of the baseline participant characteristics such as age, gender, and marathon finish time by marathon race to more closely understand our data.

Table 1: Participant Characteristics by Marathon Race (n = 10,440)

	Marathon Race					
Variable	**Boston** N = 2,021[1]	**Chicago** N = 2,256[1]	**Grandma's** N = 1,717[1]	**New York** N = 2,711[1]	**Twin Cities** N = 1,735[1]	**p-value**[2]
Age	48 (17)	48 (17)	47 (16)	51 (18)	47 (16)	<0.001
Gender						>0.9
Female	952 (47%)	1,070 (47%)	804 (47%)	1,293 (48%)	805 (46%)	
Male	1,069 (53%)	1,186 (53%)	913 (53%)	1,418 (52%)	930 (54%)	
Finish Time (s)	11,267 (2,791)	11,896 (3,775)	12,118 (3,426)	12,565 (4,620)	11,911 (3,078)	<0.001

[1] Mean (SD); n (%)

[2] Kruskal-Wallis rank sum test; Pearson's Chi-squared test

We see that the average age of participants ranges from 47 to 51 years, with a statistically significant

variation across races, indicating differences in age distribution. Gender representation is consistent, with females accounting for approximately 47-48% and males 52-54% across all races, showing no significant variation by gender. Finish times vary considerably, with mean times ranging from 11,267 seconds (~3 hours 7 minutes) in Boston to 12,565 seconds (~3 hours 29 minutes) in New York. This difference is statistically significant, highlighting variability in race performance. Overall, our dataset reveals differences in age and performance while maintaining a balanced gender distribution across the races.

Next we create a summary table looking at the characteristics of marathon runners specificlaly stratified by each individual race.

Table 2: Summary Table of Runner Characteristics

Characteristic	Race				
	Boston N = 18[1]	**Chicago** N = 20[1]	**Grandma's** N = 16[1]	**New York** N = 22[1]	**Twin Cities** N = 16[1]
Flag					
Green	7 (39%)	12 (60%)	6 (38%)	7 (32%)	7 (44%)
Red	1 (5.6%)	1 (5.0%)	2 (13%)	0 (0%)	1 (6.3%)
White	9 (50%)	6 (30%)	0 (0%)	11 (50%)	5 (31%)
Yellow	1 (5.6%)	1 (5.0%)	8 (50%)	4 (18%)	3 (19%)
Dry bulb temperature	11.6 (6.0)	12.4 (6.2)	18.9 (3.4)	11.7 (4.8)	13.2 (5.7)
Wet bulb temperature	7.6 (3.9)	8.6 (5.9)	14.9 (2.5)	7.6 (5.1)	9.9 (5.6)
Percent relative humidity	35 (35)	61 (11)	49 (36)	27 (31)	41 (35)
Black globe temperature	24 (9)	25 (6)	32 (8)	21 (6)	25 (7)
Solar radiation in Watts	654 (191)	460 (96)	679 (195)	401 (134)	437 (143)
Dew Point	3 (5)	5 (7)	12 (3)	3 (7)	6 (8)
Wind	12.0 (4.6)	8.2 (3.3)	9.2 (2.9)	11.2 (4.7)	8.8 (3.3)
WBGT	11.3 (4.6)	12.1 (5.9)	18.6 (3.3)	10.7 (5.0)	13.3 (5.6)
Air Quality Index	38 (8)	21 (7)	30 (9)	20 (5)	23 (8)

[1] n (%); Mean (SD)

From our summary table we see that flags vary significantly by race. "Green" flags are most common in Chicago (60%) and least common in New York (32%). Grandma's has the highest proportion of "Yellow" flags (50%), indicating cautionary conditions. Grandma's has the highest average dry bulb temperature (18.9°C, SD = 3.4), suggesting warmer conditions, while Boston has the lowest (11.6°C, SD = 6.0), reflecting cooler weather. Grandma's also leads with the highest wet bulb temperature (14.9°C, SD = 2.5), while Twin Cities (9.9°C, SD = 5.6) is moderate compared to Boston and New York, which have the lowest averages (7.6°C, SD = 3.9 and 5.1, respectively). Chicago experiences the highest relative humidity (61%,

SD = 11%), while New York has the lowest (27%, SD = 31%). Grandma's shows the highest black globe temperature (32°C, SD = 8), indicating exposure to higher combined radiant and ambient heat levels, while New York has the lowest (21°C, SD = 6). Grandma's also exhibits the highest average solar radiation (679 W, SD = 195), adding to the heat stress for runners. Chicago has the lowest solar radiation (460 W, SD = 96). The dew point is highest in Grandma's (12°C, SD = 3) and lowest in Boston and New York (3°C, SD = 5 and 7, respectively). Boston shows the highest wind speeds on average (12.0 km/h, SD = 4.6), while Chicago has the lowest (8.2 km/h, SD = 3.3). WBGT is highest in Grandma's (18.6°C, SD = 3.3), indicating the most challenging thermal conditions. Boston and New York have the lowest WBGT values (11.3°C and 10.7°C, respectively), suggesting milder heat stress. Lastly, Boston has the highest average air quality (38, SD = 8), while Chicago has the best air quality (AQI = 21, SD = 7).

Overall, Grandma's stands out with the most extreme heat and humidity conditions (high dry bulb, wet bulb, black globe temperatures, WBGT, and solar radiation). Twin Cities represents a race with moderate conditions across most metrics. It has higher wet bulb temperatures and WBGT than Boston or New York but avoids the extremes of Grandma's. Both Boston and New York, despite being cooler, exhibit notable differences in factors such as wind and AQI, which could still impact performance. Chicago offers moderate conditions with higher humidity but lower heat stress overall.

We next will look at the effects of age on marathon performance stratified by race, gender, and age.

Figure 1: Effects of Age on Marathon Performance by Race and Gender

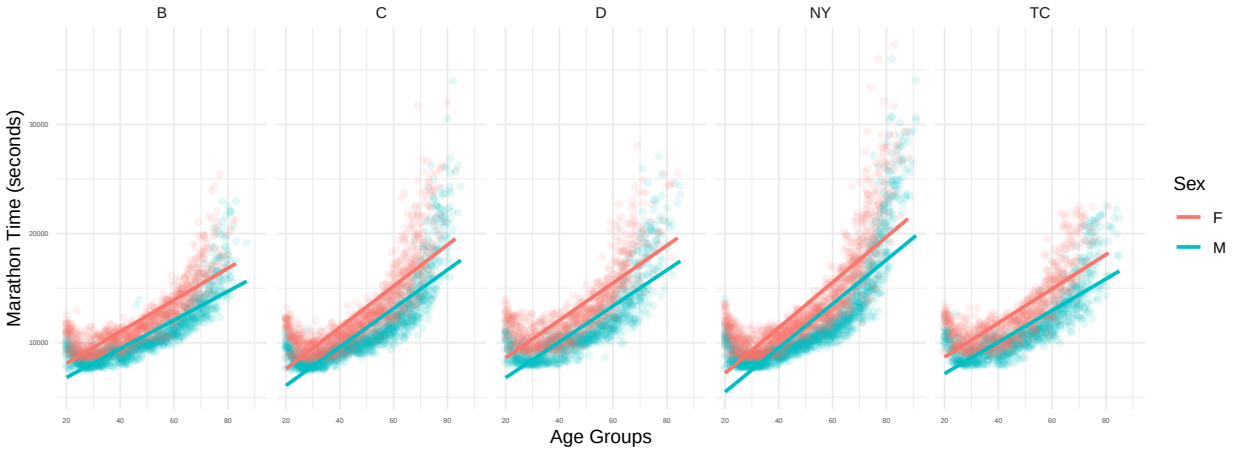


Figure 1 illustrates the effects of age on marathon performance across different race locations, with the results stratified by gender age. Each point represents individual marathon times in seconds. A general trend is visible where marathon times increase with age across both men and women, especially in the older age groups. Men tend to have slightly faster finish times across most age groups, as indicated by the clustering of blue points (men) lower than the red points (women). The trends show that men tend to outperform women across all age groups, although the difference appears more pronounced in older age brackets. Younger runners, especially those in their 20s and 30s, achieve faster marathon times, indicating that these are likely peak performance years for endurance athletes.

Next we look at the average finish time grouped by age, sex, and race.

Figure 2: Average Marathon Finish Time by Age, Sex, and Race

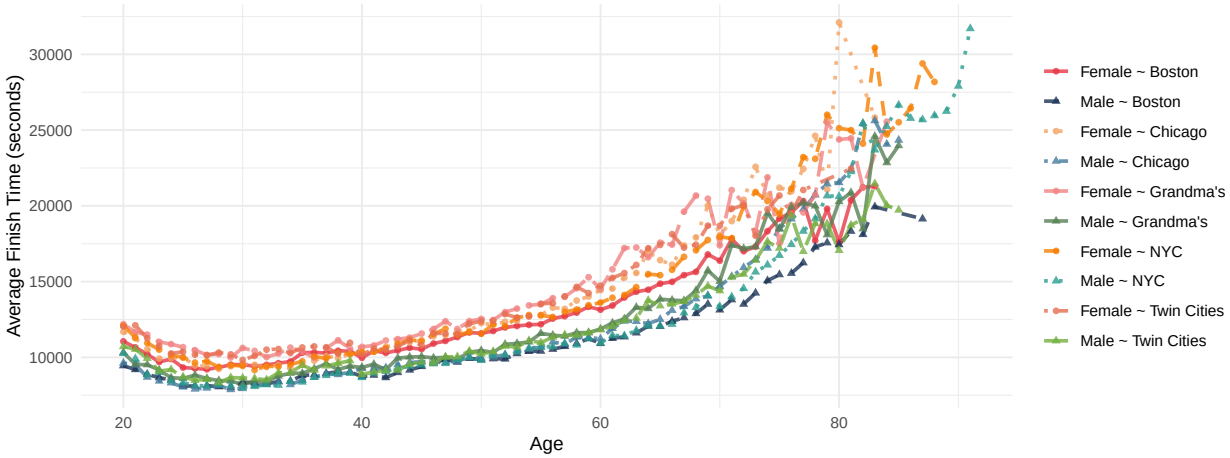
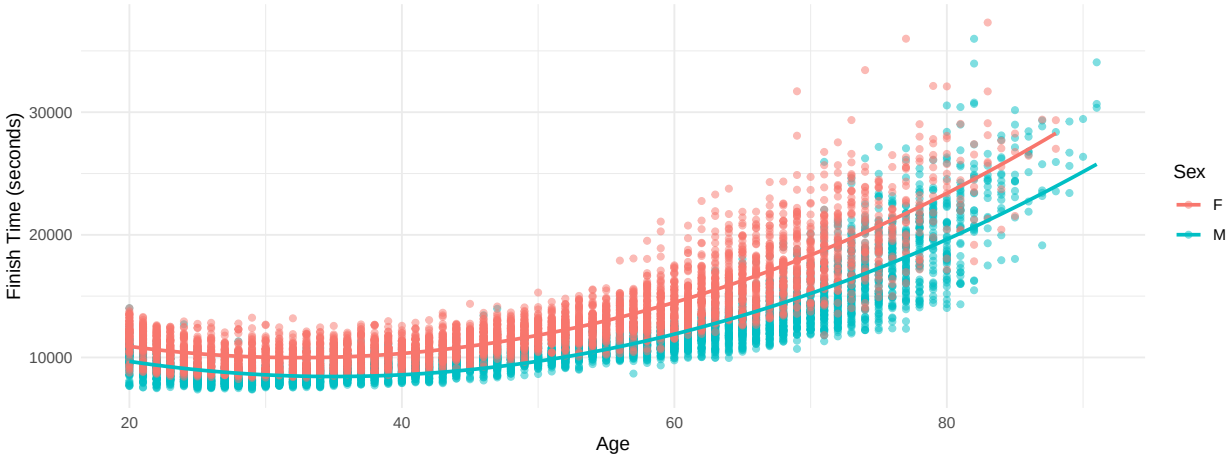


Figure 2 highlights that marathon times are relatively stable between the ages of 20 and 40, with minimal differences across race locations for both men and women. However, after age 40, a clear upward trend in finish times is visible, indicating a steady decline in performance as age increases. Interestingly, Grandma's marathon shows a more pronounced increase in finish times for both men and women, especially after age 60, compared to other races. The variability becomes more extreme after age 70, particularly for women in some races, which suggests increasing performance differences in older age groups.

We look to fit a quadratic polynomial regression model to examine how finish time changes with age for each gender separately because the relationship between age and marathon finish times is not linear and our data shows a curved trend.

Model	Term	Estimate	Std. Error	t-Value	p-Value
Men	(Intercept)	11357.34	20.222	561.629	0
Men	poly(Age, 2)1	221021.36	1501.893	147.162	0
Men	poly(Age, 2)2	120133.12	1501.893	79.988	0
Women	(Intercept)	12692.62	22.625	561.009	0
Women	poly(Age, 2)1	201455.60	1587.598	126.893	0
Women	poly(Age, 2)2	104675.43	1587.598	65.933	0

Figure 3: Marathon Finish Time vs. Age for Men and Women



Both our models and figure 3 reveal that age has a strong non-linear effect on marathon performance for both men and women, with significant linear and quadratic terms. Men's baseline finish time is around

11,357 seconds (3 hours and 9 minutes), while women's is higher at 12,692 seconds (3 hours and 31 minutes). The R-squared values indicate that age explains a large portion of the variation in marathon times for both men (83.6%) and women (80.6%). However, the quadratic model does not capture the precise peak age of performance effectively, as the parabolic shape forces a single minimum point rather than accommodating the broader range of peak performance typically observed between the ages of 25 and 35.

Figure 3 illustrates the U-shaped relationship between age and performance, with finish times lowest between ages 20-30, then increasing steadily after 30-40, and accelerating sharply after 50. To interpret the rate of slowing, the quadratic terms in Table 1 suggest that for men, finish times increase by approximately 2 minutes for each year after 40, and this rate accelerates with age. Similarly, women experience a slightly steeper increase in finish times, slowing by about 2.5 minutes per year after 40. This pattern indicates that the physiological effects of aging impact women slightly more than men.

Women consistently have slower finish times than men, and the performance decline becomes particularly steep after age 40-50 for both sexes, with greater variability seen in older runners. In summary, increasing age has a clear and substantial effect on marathon performance, with both men and women experiencing a steady decline after their early 30s. Although men tend to have faster finish times than women across all age groups, the rate of performance deterioration due to aging is significant for both sexes, and it becomes especially steep as runners enter their 50s and beyond.

AIM 2

Explore how various environmental factors influence marathon performance and examine whether these impacts vary across different age groups and genders.

For the second aim, we investigate how environmental factors influence marathon performance and whether these effects vary by gender and across age groups. To do this, we fit a multiple regression model that includes both men and women and explicitly tests for gender-specific effects by including interaction terms with `Sex`. This approach allows us to examine how environmental conditions and age affect marathon performance for men and women separately, while also testing whether these effects differ between genders.

The model includes the following environmental variables: Humidity, Solar Radiation, Wind, AQI, and Wet Bulb Globe Temperature (WBGT). WBGT is used to represent Dry Bulb Temprature, Wet Bulb Temprature, Black Globe Temprature, Humidity, and Dew Point as it represents a weighted average of dry bulb, wet bulb, and globe temperatures, providing a comprehensive measure of heat stress. In this way, we can avoid multicollinearity as Wet Bulb Globe Temprature is a composite variable hat already incorporates aspects of these other characteristics (see correlation matrix under Aim 3 for further explanation). Our dependent variable will be the percent off current course record for gender to standardize times.

Given our previous findings, to account for the non-linear relationship between age and marathon performance, we include a quadratic term for Age (`Age^2`). Interaction terms between environmental factors, the quadratic polynomial of Age, and `Sex` allow us to test how the effects of environmental conditions vary with age and gender.

In our initial model, we included a comprehensive set of predictors, including multiple higher-order interaction terms between environmental factors, age, and sex. While this model provided a strong fit, with an adjusted R-squared of 0.8406, it exhibited several drawbacks. The inclusion of numerous three-way interactions introduced unnecessary complexity, making the model challenging to interpret and increasing the risk of overfitting. Additionally, many of these higher-order terms were not statistically significant, contributing little to the explanatory power of the model while potentially inflating multicollinearity and standard errors.

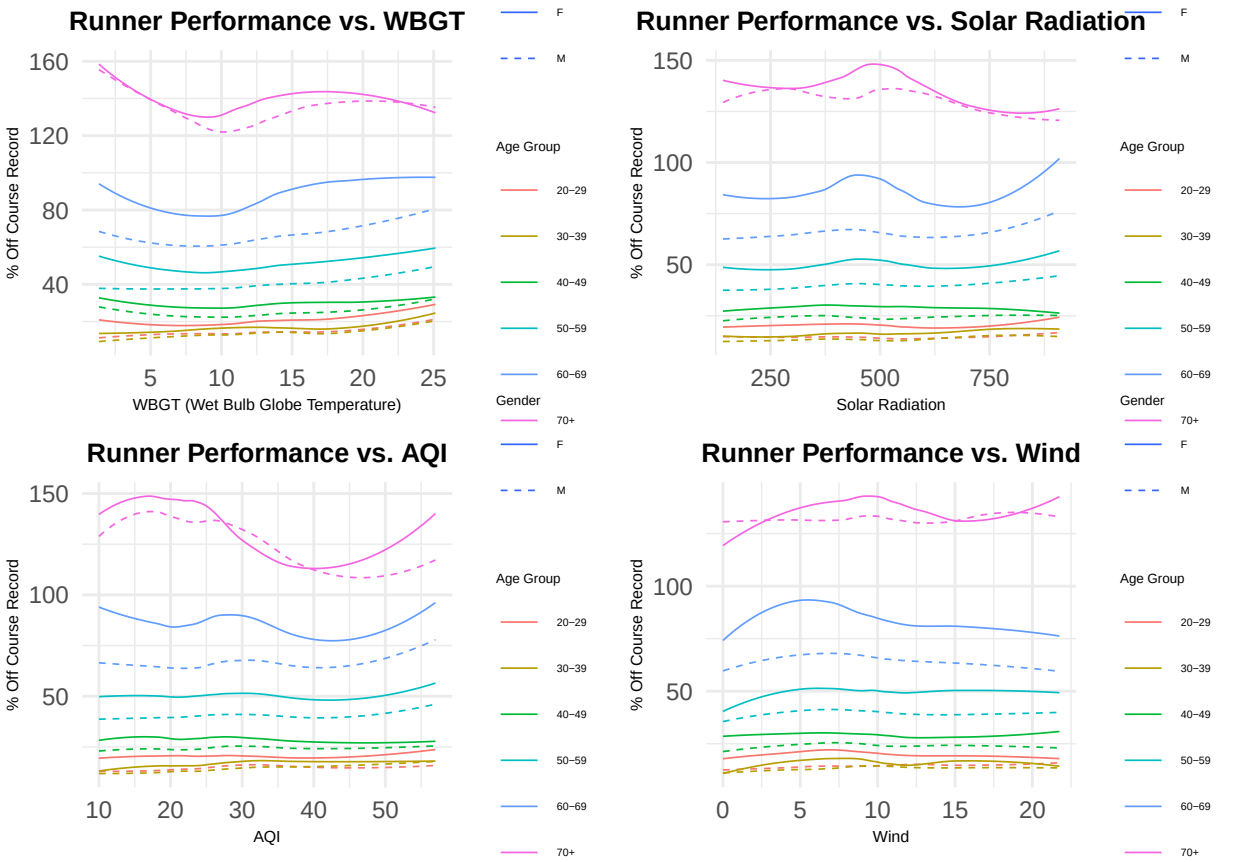
To address these issues, we refined the model by systematically removing non-significant interaction terms, resulting in a simplified model which is displayed below.

Model	Term	Estimate	Std. Error	t-Value
(Intercept)	48.218	0.910	52.986	0.000
Solar_Radiation	0.002	0.001	1.947	0.052
poly(Age, 2)1	4119.899	96.290	42.786	0.000
poly(Age, 2)2	2161.334	97.599	22.145	0.000
Wind	0.138	0.049	2.801	0.005
aqi	-0.199	0.023	-8.795	0.000
Wet_Bulb_Globe_Temp	0.610	0.051	12.010	0.000
SexM	-8.786	0.907	-9.692	0.000
Solar_Radiation:poly(Age, 2)1	0.311	0.117	2.663	0.008
Solar_Radiation:poly(Age, 2)2	-0.062	0.117	-0.528	0.598
poly(Age, 2)1:Wind	8.569	4.973	1.723	0.085
poly(Age, 2)2:Wind	15.417	4.915	3.137	0.002
poly(Age, 2)1:aqi	-27.996	2.303	-12.154	0.000
poly(Age, 2)2:aqi	-21.312	2.332	-9.140	0.000
poly(Age, 2)1:Wet_Bulb_Globe_Temp	30.656	5.587	5.487	0.000
poly(Age, 2)2:Wet_Bulb_Globe_Temp	19.778	5.693	3.474	0.001
Wet_Bulb_Globe_Temp:SexM	-0.072	0.065	-1.105	0.269
poly(Age, 2)1:SexM	-383.734	94.263	-4.071	0.000
poly(Age, 2)2:SexM	416.140	94.678	4.395	0.000
poly(Age, 2)1:Wet_Bulb_Globe_Temp:SexM	-7.757	6.852	-1.132	0.258
poly(Age, 2)2:Wet_Bulb_Globe_Temp:SexM	-27.624	6.915	-3.995	0.000

Our final model provides significant insights into how age, gender, and environmental factors impact marathon performance. Both the linear and quadratic components of age are highly significant, confirming that age is a key driver of marathon finish times, with a sharp decline in performance for older runners. Gender differences are also substantial, with men generally achieving faster finish times than women, as shown by the statistically significant coefficient for `SexM`.

Environmental factors play an essential role in influencing marathon performance, with significant contributions from WBGT, AQI, and wind speed. Among these, WBGT demonstrates the strongest effect, with higher temperatures slowing performance substantially, emphasizing the importance of heat stress in endurance events. AQI also has a negative effect, with poorer air quality leading to slower finish times, emphasizing the sensitivity of marathon runners to atmospheric conditions. Wind speed positively impacts performance, suggesting that tailwinds or lower wind resistance may improve outcomes for runners.

The model also highlights significant interaction effects, revealing nuanced relationships between environmental conditions, age, and gender. For instance, older runners are more negatively impacted by higher temperatures and poorer air quality. Furthermore, the interaction between WBGT and `Sex` indicates that men and women respond differently to heat stress, with men being slightly less affected than women. These interaction terms emphasize that environmental conditions disproportionately affect specific groups, such as older runners and women. Our final model explains approximately 66% of the variability in marathon finish times (R-squared = 0.8403), demonstrating a strong overall fit while avoiding unnecessary complexity.



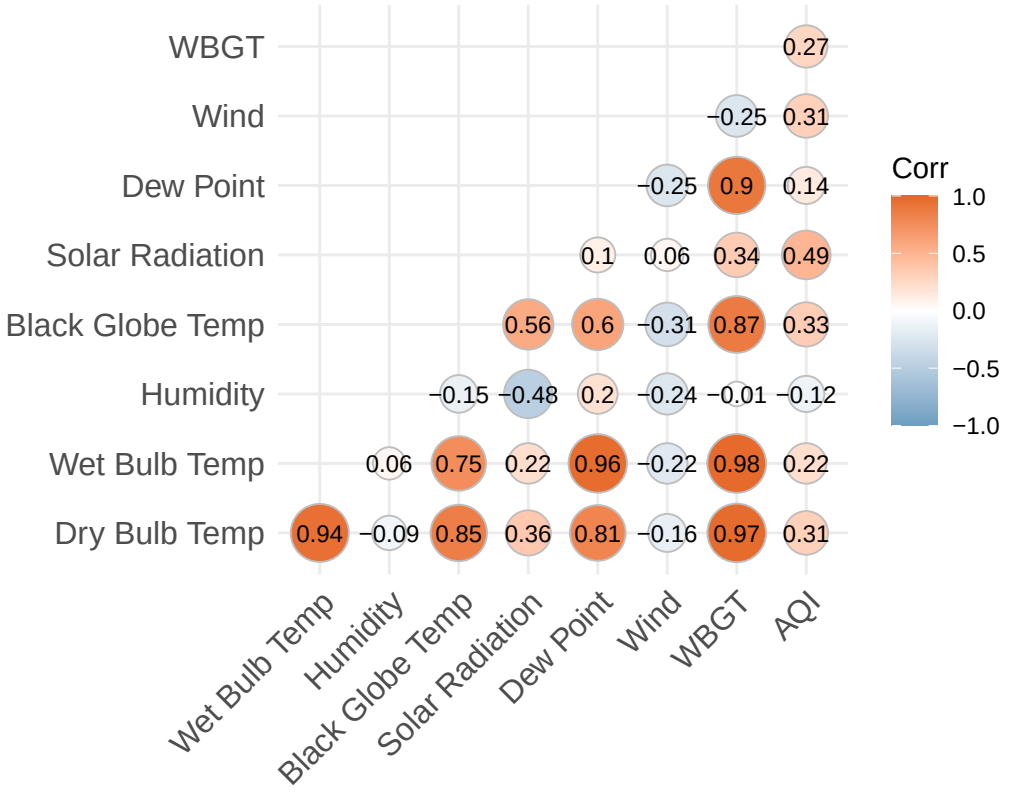
Across all environmental factors, female runners consistently outperform male runners, as seen in the lower percentage off course record values. This trend is particularly evident in the WBGT and Solar Radiation plots. Similar to what we've found before, in all the figures we see that older age groups tend to show higher percentages off the course record, indicating a decline in performance with age. Simultaneously, younger groups exhibit more consistent and lower percent off course record values. We see a higher percentage off course record as WBGT increases, with the steepest increases for older runners and a noticeable difference between genders. There's a slight upward trend in percentage off course record for most groups as solar radiation increases, indicating that higher radiation levels impair performance. As for AQI, performance appears relatively stable with minimal changes for younger age groups. Older age groups show slightly higher percentage off course record values, but the trends are not as pronounced. Lastly, performance tends to worsen slightly with increasing wind speeds. The trend is most noticeable in older age groups, while younger age groups remain less affected.

AIM 3

Identify which weather parameters have the largest effect on marathon performance, hypothesizing that increased environmental temperatures and adverse weather conditions negatively impact overall performance.

To address the hypothesis that increased environmental temperatures and adverse weather conditions negatively impact overall marathon performance, we can use regression modeling and correlation analysis to identify which weather parameters have the largest impact on finish times.

Figure 5: Correlation Between Environmental Parameter



Our correlation heatmap includes strong positive correlations between Dry Bulb Temperature, Wet Bulb Temperature, and WBGT, which is expected as these metrics are interrelated and often combine to reflect overall heat stress conditions. For example, the high correlation between Dry Bulb Temp and Wet Bulb Temp (0.94) suggests a strong dependency between these two measures of temperature and humidity. Another notable relationship is between Dew Point and Wet Bulb Temp (0.96), as both metrics capture aspects of atmospheric moisture. Similarly, Black Globe Temperature shows a strong correlation with Dry Bulb Temperature (0.85), likely because both are influenced by ambient temperature. Conversely, Humidity exhibits weaker and somewhat negative correlations with parameters like Solar Radiation (-0.48), indicating that high humidity is often associated with lower levels of solar exposure, potentially due to cloud cover. Interestingly, Wind has a weak negative correlation with Dry Bulb Temp (-0.16) and WBGT (-0.22), which might reflect how increased wind can provide a cooling effect, mitigating some aspects of heat stress.

Overall, the heatmap highlights how temperature, humidity, and solar radiation interact to influence environmental conditions, with WBGT standing out as a composite metric effectively summarizing these factors. It is for these reasons why we only included WBGT and excluded other variables such as Dry Bulb Temprature, Wet Bulb Temprature, Black Globe Temprature, Humidity, and Dew Point in our regression analysis from Aim 2.

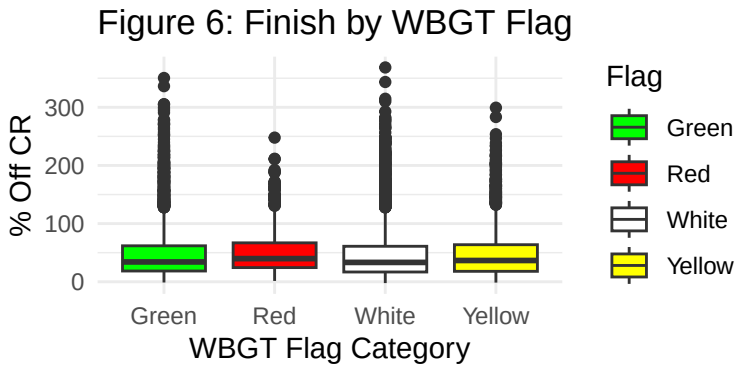
From our multiple regression analysis in Aim 2, we can conclude that Wet Bulb Globe Temperature emerges as the most impactful factor, with higher values leading to slower performance. This measure, which combines temperature, humidity, and solar radiation, effectively captures the heat stress conditions that negatively influence runners. Air Quality Index also plays a critical role, with poorer air quality strongly linked to slower performance. Additionally, wind speed positively impacts performance, likely due to its cooling effect, as shown by its statistically significant positive relationship with marathon times.

While solar radiation shows a smaller direct effect on performance, its interaction with age suggests that its impact varies across age groups, with younger runners potentially adapting better to higher radiation

levels. Furthermore, the polynomial terms for age and interactions with WBGT and other weather parameters indicate that older runners are more adversely affected by heat and adverse environmental conditions. These results suggest that WBGT, AQI, and wind speed are the most critical factors influencing marathon performance, with demographic differences further shaping the extent of these effects.

Next, we explore the Flag parameter which categorizes Wet Bulb Globe Temperature (WBGT) based on heat illness risk levels. We want to show how marathon performance is impacted by the different flag levels below:

- White: WBGT < 10°C
- Green: WBGT 10-18°C
- Yellow: WBGT 18-23°C
- Red: WBGT 23-28°C
- Black: WBGT > 28°C



We see from the figure above that runners in the Green flag category have the most consistent and favorable percentages off the course record, suggesting optimal running conditions. In contrast, Red flag conditions show a slightly higher median percentage off course record, though the distribution is less variable, indicating moderately challenging conditions. The White flag and Yellow flag categories exhibit more variability and numerous outliers, suggesting that some runners struggle in these conditions. Overall, as the WBGT flag moves from more favorable (Green) to more challenging (Red/Yellow), there is an observable trend towards longer and more varied percentages off the course record, highlighting the impact of environmental conditions on marathon performance. Next we use an ANOVA test to assess whether the differences between the WBGT flag categories are actually statistically significant.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Flag	3	17840.95	5946.983	2.875116	0.0347595
Residuals	10436	21586161.79	2068.433	NA	NA

Since the p-value is less than 0.05, we can conclude that the percentages off the course record vary significantly across the different WBGT flag categories. To determine which specific flag categories differ from one another, we can run a Tukey's HSD test. This will allow us to perform pairwise comparisons between the flag categories to identify where the significant differences lie.

	diff	lwr	upr	p adj
Red-Green	4.7114239	-0.5847787	10.0076264	0.1013821
White-Green	-0.6027354	-3.2295674	2.0240967	0.9352698
Yellow-Green	1.7398631	-1.4728482	4.9525744	0.5046833
White-Red	-5.3141592	-10.6789183	0.0505998	0.0533030
Yellow-Red	-2.9715608	-8.6462331	2.7031116	0.5338535
Yellow-White	2.3425985	-0.9819153	5.6671122	0.2683138

Although the overall ANOVA showed significant differences in percentages off the course records across the WBGT flag categories, the pairwise comparisons do not show any specific flag categories having significant differences after adjusting for multiple comparisons. The most notable result is the White vs Red comparison, which is close to significance and suggests that the percentages off the course records might be slower in Red flag conditions (higher WBGT) compared to White (lower WBGT).

Conclusion

This analysis provided valuable insights into the various impacts of age, gender, and environmental conditions on marathon performance. The results confirmed a clear non-linear relationship between age and marathon performance, with peak performance occurring in the 20s and 30s, followed by a steady decline after the age of 40. Men generally outperformed women across all age groups, but the rate of performance deterioration with age was similar for both genders. Older runners displayed greater resilience in maintaining a consistent pace compared to younger runners, confirming previous research.

Environmental factors emerged as critical determinants of performance. Wet Bulb Globe Temperature was identified as the most significant variable, with higher temperatures and humidity having a pronounced negative effect, particularly on older runners and women. Poor air quality further compounded performance challenges, emphasizing the sensitivity of endurance athletes to atmospheric conditions. Conversely, wind was found to positively influence performance, likely due to its cooling effects.

The analysis of WBGT flag categories reinforced the significance of environmental stressors, with less favorable categories correlating with poorer performance and higher variability. While this finding emphasizes the challenges of running in adverse weather, it also highlights the importance of optimizing race conditions to ensure runner safety and performance.

Overall, this study highlights the intricatces of physiological and environmental factors in shaping marathon outcomes. These insights can inform targeted strategies for training, race planning, and public health recommendations, ensuring that runners across age and gender can achieve their optimal performance while mitigating environmental risks. Future research could expand on this foundation by exploring additional demographic variables, such as training levels or acclimatization, to further refine our understanding of endurance performance dynamics.

References

Deaner, R. O., Addona, V., Hanley, B., & Carter, R. E. (2015). Men are more likely than women to slow in the marathon. Medicine & Science in Sports & Exercise, 47(3), 607–616. https://doi.org/10.1249/MSS.0000000000000432

Nikolaidis, P. T. (2019). Pace variability among age groups in marathon races. Journal of Strength and Conditioning Research, 33(6), 1616–1623. https://doi.org/10.1519/JSC.0000000000002467

Knechtle, B., Valeri, F., Zingg, M. A., Rosemann, T., & Rüst, C. A. (2021). The influence of temperature and humidity on marathon performance of older age groups in the New York City Marathon. Research in Sports Medicine, 29(1), 17–32. https://doi.org/10.1080/15438627.2020.1719240

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
# set wd, create DFs, load in proper packages
setwd("/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1")
final_marathon_df <- read.csv("final_marathon_df.csv")
library(gtsummary)
library(dplyr)
library(gt)
library(tidyr)
library(tidyverse)
library(HDSinRdata)
library(ggplot2)
library(readr)
library(reshape2)
library(car)
library(kableExtra)
library(broom)
library(corrplot)
library(gridExtra)
library(ggcorrplot)
# change column names (marathon_df)
marathon_df <- marathon_df %>%
  rename(Race = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
         Sex = Sex..0.F..1.M.,
         Age = Age..yr.,
         Finish = X.CR,
         Dry_Bulb_Temp = Td..C,
         Wet_Bulb_Temp = Tw..C,
         Humidity = X.rh,
         Black_Globe = Tg..C,
         Solar_Radiation = SR.W.m2,
         Dew_Point = DP,
         Wet_Bulb_Globe_Temp = WBGT
         )
# change race/sex variable names
marathon_df <- marathon_df %>%
  mutate(Race = case_when(Race == "0" ~ "B",
                          Race == "1" ~ "C",
                          Race == "2" ~ "NY",
                          Race == "3" ~ "TC",
                          Race == "4" ~ "D")) %>%
  mutate(Sex = case_when(Sex == "0" ~ "F",
                         Sex == "1" ~ "M"))

course_record <- course_record %>%
  rename(Sex = Gender)
# join marathon_df and course_record
marathon_time <- left_join(marathon_df, course_record,
                           by=c("Race", "Sex", "Year"))
# write new CSV for marathon_time
write_csv(marathon_time, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1/ma
marathon_time <- read.csv("marathon_time.csv")
```

```r
# create actual time column
# convert "HH:MM:SS" to seconds
time_to_seconds <- function(time_str) {
  parts <- as.numeric(unlist(strsplit(time_str, ":")))
  return(parts[1] * 3600 + parts[2] * 60 + parts[3])
}

# apply to convert each CR to seconds
marathon_time$CR_seconds <- sapply(marathon_time$CR, time_to_seconds)

# calculate actual race time in seconds based on Finish percentage off the CR time
marathon_time$Actual_Time_Seconds <- marathon_time$CR_seconds * (1 + (marathon_time$Finish / 100))
marathon_df <- marathon_time %>%
  rename(Finish_seconds = Actual_Time_Seconds)
# write new CSV for marathon_df with actual course times
write_csv(marathon_df, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1/mara
marathon_df <- read.csv("marathon_df.csv")
# missing data by Race
missing_tb2 <- marathon_df %>%
  group_by(Race, Year) %>%
  summarise(
    Flag_Missing = sum(is.na(Flag)),
    Dry_Bulb_Temp_Missing = sum(is.na(Dry_Bulb_Temp)),
    Wet_Bulb_Temp_Missing = sum(is.na(Wet_Bulb_Temp)),
    Humidity_Missing = sum(is.na(Humidity)),
    Black_Globe_Missing = sum(is.na(Black_Globe)),
    Solar_Radiation_Missing = sum(is.na(Solar_Radiation)),
    Dew_Point_Missing = sum(is.na(Dew_Point)),
    Wind_Missing = sum(is.na(Wind)),
    Wet_Bulb_Globe_Temp_Missing = sum(is.na(Wet_Bulb_Globe_Temp))
  )

missing_tb2 <- as.data.frame(t(missing_tb2))
colnames(missing_tb2) <- missing_tb2[1, ]
missing_tb2 <- missing_tb2[-1, ]
missing_tb2$Variable <- factor(row.names(missing_tb2))
missing_tb2 <- missing_tb2[, c(ncol(missing_tb2), 1:(ncol(missing_tb2) - 1))]

missing_tb2 <- missing_tb2 %>%
  mutate(Variable = case_when(
    Variable == "Flag_Missing" ~ "Flag Missing",
    Variable == "Dry_Bulb_Temp_Missing" ~ "Dry Bulb Temperature Missing",
    Variable == "Wet_Bulb_Temp_Missing" ~ "Wet Bulb Temperature Missing",
    Variable == "Humidity_Missing" ~ "Humidity Missing",
    Variable == "Black_Globe_Missing" ~ "Black Globe Temperature Missing",
    Variable == "Solar_Radiation_Missing" ~ "Solar Radiation Missing",
    Variable == "Dew_Point_Missing" ~ "Dew Point Missing",
    Variable == "Wind_Missing" ~ "Wind Missing",
    Variable == "Wet_Bulb_Globe_Temp_Missing" ~ "Wet Bulb Globe Temperature Missing"
  ))

gt_table <- gt(missing_tb2) %>%
  tab_header(
```

```
    title = "Table 1: Missing Data by Race",
    subtitle = "Number of missing observations per variable stratified by race"
  ) %>%
  tab_spanner(
    label = "Race Categories",
    columns = colnames(missing_tb2)[-1]
  )

gt_table
# total percent of missing values in the dataset
total_missing <- sum(is.na(marathon_df))
total_values <- prod(dim(marathon_df))
percentage_missing <- (total_missing / total_values) * 100

# percentage of missing data is negligible since <5%, Use observed data only
# y-axis average finish, x-axis age, color F/M
#group by into age groups, Under 19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+
# Under 19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70+, typical age grou
final_marathon <- marathon_df %>%
  mutate(
    age_group = case_when(
      Age <= 19 ~ "19 and Under",
      Age > 19 & Age <= 29 ~ "20-29",
      Age > 29 & Age <= 39 ~ "30-39",
      Age > 39 & Age <= 49 ~ "40-49",
      Age > 49 & Age <= 59 ~ "50-59",
      Age > 59 & Age <= 69 ~ "60-69",
      Age > 69 ~ "70+"
    ),
    age_group = factor(
      age_group,
      level = c("Under 19", "20-29","30-39", "40-49", "50-59", "60-69", "70+")
    )
  )

 final_marathon <- na.omit(final_marathon)
aqi_values <- aqi_df %>%
   rename(Race = marathon) %>%
     mutate(
     Race = case_when(
     Race == "NYC" ~ "NY",
     Race == "Grandmas" ~ "D",
     Race == "Boston" ~ "B",
     Race == "Twin Cities" ~"TC",
     Race == "Chicago" ~ "C"
  ),
    date = as.Date(date_local, format = "%Y-%m-%d"),
    Year = as.numeric(format(date, "%Y"))
      ) %>%
   select(-date_local)

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()

course_record_try <-  final_marathon %>%
  left_join(avg_ppm, by = c("Race", "Year"))
write_csv(course_record_try, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project
FINAL_df <- read.csv("FINAL_df.csv")

# multiply avg_ppm by 1000, round to nearest whole number, and rename the column to 'aqi'
last_df <- FINAL_df %>%
  mutate(aqi = round(avg_ppm * 1000)) %>%
  select(-avg_ppm) # remove the original avg_ppm column
# write final CSV
write_csv(last_df, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1/final_ma
# TABLE 1
# n = 10, 440
summary_table <- final_marathon_df %>%
  mutate(
    Sex = ifelse(Sex == "M", "Male", "Female"),
    Race = case_when(
      Race == "B" ~ "Boston",
      Race == "C" ~ "Chicago",
      Race == "NY" ~ "New York",
      Race == "TC" ~ "Twin Cities",
      Race == "D" ~ "Grandma's"
    )
  ) %>%
  tbl_summary(
    include = c(Age, Sex, Finish_seconds),
    by = Race,
    label = list(
      Age = "Age",
      Sex = "Gender",
      Finish_seconds = "Finish Time (s)"
    ),
    missing = "no",
    statistic = all_continuous() ~ "{mean} ({sd})"
  ) %>%
  add_p() %>%
  bold_labels() %>%
  modify_header(label = "**Variable**") %>%
  modify_spanning_header(all_stat_cols() ~ "Marathon Race") %>%
  modify_caption(" Table 1: Participant Characteristics by Marathon Race (n = 10,440)")
# could not get the table to fit on the page all the way, so I took a screenshot of it in R markdown an
sum_tbl_2 <- final_marathon_df %>%
  group_by(Race, Year, Flag) %>%
  summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%
  select(Race, Year, Flag, Dry_Bulb_Temp, Wet_Bulb_Temp, Humidity, Black_Globe,
         Solar_Radiation, Dew_Point, Wind, Wet_Bulb_Globe_Temp, aqi)
# summary table 2
summary_table2 <- sum_tbl_2 %>%
```

```
  mutate(
    Race = case_when(
      Race == "B" ~ "Boston",
      Race == "C" ~ "Chicago",
      Race == "NY" ~ "New York",
      Race == "TC" ~ "Twin Cities",
      Race == "D" ~ "Grandma's"
    )) %>%
  tbl_summary(
    by = Race,
    include = c(Flag, Dry_Bulb_Temp, Wet_Bulb_Temp, Humidity, Black_Globe,
                Solar_Radiation, Dew_Point, Wind, Wet_Bulb_Globe_Temp, aqi),
    label = list(
      Flag = "Flag",
      Dry_Bulb_Temp = "Dry bulb temperature",
      Wet_Bulb_Temp = "Wet bulb temperature",
      Humidity = "Percent relative humidity",
      Black_Globe = "Black globe temperature",
      Solar_Radiation = "Solar radiation in Watts",
      Dew_Point = "Dew Point",
      Wind = "Wind",
      Wet_Bulb_Globe_Temp = "WBGT",
      aqi = "Air Quality Index"
    ),
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    missing = "no"
  ) %>%
  bold_labels() %>%
  modify_header(label = "**Characteristic**") %>%
  modify_spanning_header(all_stat_cols() ~ "**Race**") %>%
  modify_caption("**Table 2: Summary Table of Runner Characteristics**") %>%
  as_gt()

summary_table2
ggplot(final_marathon_df, aes(x = Age, y = Finish_seconds, color = Sex)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method=lm, se = FALSE) +
  labs( title = "Figure 1: Effects of Age on Marathon Performance by Race and Gender",
        x = "Age Groups",
        y = "Marathon Time (seconds)") +
  facet_grid(~Race) +
  theme_minimal()+
  theme(axis.text = element_text(size = 4))
avg_finish_time <- aggregate(Finish_seconds ~ Age + Sex + Race, data = final_marathon_df, FUN = mean)


warm_colors <- c("#E63946", "#F4A261", "#F77F00", "#E76F51", "#F28482")  # warm tones for females
cool_colors <- c("#1D3557", "#457B9D", "#2A9D8F", "#76B041", "#588157")  # cool tones for males

races <- unique(final_marathon_df$Race)
```

```r
color_mapping <- setNames(
  c(warm_colors, cool_colors),
  c(paste0("F.", races), paste0("M.", races))
)

custom_labels <- c(
  "Female ~ Boston", "Male ~ Boston", "Female ~ Chicago", "Male ~ Chicago",
  "Female ~ Grandma's", "Male ~ Grandma's", "Female ~ NYC", "Male ~ NYC",
  "Female ~ Twin Cities", "Male ~ Twin Cities"
)

line_styles <- c("solid", "dashed", "dotted", "dotdash", "longdash")  # different styles for races

ggplot(avg_finish_time, aes(x = Age, y = Finish_seconds, color = interaction(Sex, Race),
                            shape = interaction(Sex, Race), linetype = interaction(Sex, Race))) +
  geom_line(linewidth = 1, alpha = 0.8) +  # thicker and slightly transparent lines
  geom_point(linewidth = 2, alpha = 0.8) +  # points with moderate size and transparency
  scale_color_manual(values = color_mapping, labels = custom_labels) +
  scale_shape_manual(values = c(16, 17, 16, 17, 16, 17, 16, 17, 16, 17), labels = custom_labels) +
  scale_linetype_manual(values = rep(line_styles, 2), labels = custom_labels) +
  labs(title = 'Figure 2: Average Marathon Finish Time by Age, Sex, and Race',
       x = 'Age',
       y = 'Average Finish Time (seconds)') +
  theme_minimal() +
  theme(legend.title = element_blank(), legend.position = "right")

data_men <- final_marathon_df %>% filter(Sex == "M")
# quadratic polynomial model for men
model_men <- lm(Finish_seconds ~ poly(Age, 2), data = data_men)

data_women <- final_marathon_df %>% filter(Sex == "F")
# quadratic polynomial model for women
model_women <- lm(Finish_seconds ~ poly(Age, 2), data = data_women)

tidy_men <- tidy(model_men) %>% mutate(Model = "Men")
tidy_women <- tidy(model_women) %>% mutate(Model = "Women")

# combine the data
combined_results <- bind_rows(tidy_men, tidy_women) %>%
  select(Model, term, estimate, std.error, statistic, p.value)

colnames(combined_results) <- c("Model", "Term", "Estimate", "Std. Error", "t-Value", "p-Value")

# table
kable(combined_results, format = "latex", booktabs = TRUE, digits = 3) %>%
  kable_styling(latex_options = c("HOLD_position"))

# plot data for both men and women
ggplot(final_marathon_df, aes(x = Age, y = Finish_seconds, color = Sex)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
  labs(title = "Figure 3: Marathon Finish Time vs. Age for Men and Women",
       x = "Age", y = "Finish Time (seconds)") +
```

```r
  theme_minimal()
# just use Wet Bulb Globe Temperature since its Weighted average of dry bulb, wet bulb, and globe tempe
# include the quadratic term for Age^2 in the regression models (since it previously improved R2)

# model to test gender-specific effects
model_combined <- lm(Finish ~ Solar_Radiation * poly(Age, 2) * Sex +
                       Wind * poly(Age, 2) * Sex +
                       aqi * poly(Age, 2) * Sex +
                       Wet_Bulb_Globe_Temp * poly(Age, 2) * Sex,
                     data = final_marathon_df)
# refined model after removing non-significant terms
model_refined <- lm(Finish ~ Solar_Radiation * poly(Age, 2) +
                       Wind * poly(Age, 2) +
                       aqi * poly(Age, 2) +
                       Wet_Bulb_Globe_Temp * poly(Age, 2) * Sex,
                     data = final_marathon_df)

model_refined<- tidy(model_refined)

colnames(model_refined) <- c("Model", "Term", "Estimate", "Std. Error", "t-Value", "p-Value")

# table
kable(model_refined, format = "latex", booktabs = TRUE, digits = 3) %>%
  kable_styling(latex_options = c("HOLD_position"))
# Create the plot
plot1 <- ggplot(final_marathon_df, aes(x = Wet_Bulb_Globe_Temp, y = Finish, color = age_group, linetype
  geom_smooth(se = FALSE, size = .3) +
  scale_linetype_manual(values = c("M" = "dashed", "F" = "solid")) +
  labs(
    title = "Runner Performance vs. WBGT",
    x = "WBGT (Wet Bulb Globe Temperature)",
    y = "% Off Course Record",
    color = "Age Group",
    linetype = "Gender"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.title = element_text(size = 6),
    legend.title = element_text(size = 5),
    legend.text = element_text(size = 4)
  )

plot2 <- ggplot(final_marathon_df, aes(x = Solar_Radiation, y = Finish, color = age_group, linetype = S
  geom_smooth(se = FALSE, size = .3) +
  scale_linetype_manual(values = c("M" = "dashed", "F" = "solid")) +
  labs(
    title = "Runner Performance vs. Solar Radiation",
    x = "Solar Radiation",
    y = "% Off Course Record",
    color = "Age Group",
    linetype = "Gender"
  ) +
```

```
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.title = element_text(size = 6),
    legend.title = element_text(size = 5),
    legend.text = element_text(size = 4)
  )

plot3 <- ggplot(final_marathon_df, aes(x = aqi, y = Finish, color = age_group, linetype = Sex)) +
  geom_smooth(se = FALSE, size = .3) +
  scale_linetype_manual(values = c("M" = "dashed", "F" = "solid")) +
  labs(
    title = "Runner Performance vs. AQI",
    x = "AQI",
    y = "% Off Course Record",
    color = "Age Group",
    linetype = "Gender"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.title = element_text(size = 6),
    legend.title = element_text(size = 5),
    legend.text = element_text(size = 4)
  )

plot4 <- ggplot(final_marathon_df, aes(x = Wind, y = Finish, color = age_group, linetype = Sex)) +
  geom_smooth(se = FALSE, size = .3) +
  scale_linetype_manual(values = c("M" = "dashed", "F" = "solid")) +
  labs(
    title = "Runner Performance vs. Wind",
    x = "Wind",
    y = "% Off Course Record",
    color = "Age Group",
    linetype = "Gender"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.title = element_text(size = 6),
    legend.title = element_text(size = 5),
    legend.text = element_text(size = 4)
  )

grid.arrange(plot1, plot2, plot3, plot4, nrow = 2)
# correlation matrix
filtered_data <- final_marathon_df %>%
  select(-c(Finish_seconds, CR_seconds, Finish, Age, Year))
numeric_columns <- filtered_data %>% select(where(is.numeric))
correlation_matrix <- cor(numeric_columns, use = "complete.obs")

colnames(correlation_matrix) <- c("Dry Bulb Temp", "Wet Bulb Temp", "Humidity", "Black Globe Temp", "Sol
rownames(correlation_matrix) <- c("Dry Bulb Temp", "Wet Bulb Temp", "Humidity", "Black Globe Temp", "Sol
```

```r
ggcorrplot(correlation_matrix,
           method = "circle",
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           colors = c("#6D9EC1", "white", "#E46726"),
           title = "Figure 5: Correlation Between Environmental Parameters",
           ggtheme = theme_minimal())
# Finish Times by WBGT Flag Category
ggplot(final_marathon_df, aes(x = Flag, y = Finish, fill = Flag)) +
  geom_boxplot() +
  labs(title = "Figure 6: Finish by WBGT Flag",
       x = "WBGT Flag Category", y = "% Off CR") +
  theme_minimal() +
  scale_fill_manual(values = c("White" = "white", "Green" = "green", "Yellow" = "yellow",
                               "Red" = "red", "Black" = "black"))

# ANOVA test
anova_result <- aov(Finish ~ Flag, data = final_marathon_df)
anova_summary <- summary(anova_result)

kable(anova_summary[[1]])
# Tukey's HSD test
tukey_result <- TukeyHSD(anova_result)

kable(as.data.frame(tukey_result[[1]]))
```