Exploring the Effect of Weather on Marathon Performances

Yingqiu Huang

Abstract

Weather conditions are crucial in marathon races and have a significant impact on runners' performances. The purpose of this report is to conduct Exploratory data analysis (EDA) to investigate the relationship between weather conditions and marathon performances, and how this impact differs across age and gender. EDA plots showed no obvious differences in marathon performance across different weather conditions. Yet the statistical model suggests that some weather parameters have a statistically significant impact on performances. However, the estimated effect sizes for these variables are relatively small, indicating that although the associations are statistically significant, their practical impact on marathon performance might be negligible. Limitations of this study include different runners each year & marathon; small sample size for senior runners and air quality measures are not clearly documented, suggesting the need for further research to develop a more accurate result.

Introduction

Weather conditions significantly influence marathon performance, with heat and humidity playing critical roles. Previous studies have established that increasing Wet Bulb Globe Temperature (WBGT) leads to a progressive slowing of marathon performances across runners of varying abilities, with the impact being more pronounced in male runners than in female runners (Ely et al., 2007). Beyond marathon running, research on aerobic performance in controlled environments highlights that heat stress degrades physical performance even without excessive hyperthermia, as pacing strategies shift and total work output diminishes in hotter conditions (Ely et al., 2010). This finding suggests that environmental factors, even in modestly elevated temperatures, can disrupt optimal pacing in marathon races. Age and sex differences further compound the complexity of understanding marathon performance under varying weather conditions. Aging influences thermoregulation, including altered sweating and cardiovascular responses to heat stress (Kenney and Munce, 2003). Moreover, sex differences in thermoregulation and endurance running biomechanics include females' greater carbohydrate preservation and even pacing, while males typically exhibit higher VO2max and muscle power (Besson et al., 2022).

Given the complicated nature of marathon performance, variables such as WBGT, solar radiation, air quality, wind, and humidity are essential to assessing weather conditions comprehensively. This report leverages data from five major marathon races (Boston, Chicago, New York City, Grandma's, and Twin Cities) spanning 15–20 years to explore the relationship between weather conditions and runner performance, and how that relationship differs under different sex and age conditions. Analytical methods include data quality checks, missing data analysis, summary tables and plots, correlation inspection, and regression modeling.

Data Collection

This report utilizes four datasets to analyze the impact of weather and environmental factors on marathon performance. The main dataset contains top single-age performances from five major marathons (Boston, Chicago, NYC, Grandma's, and Twin Cities) over a period of 15–20 years, covering runners aged 14 to 85, in total has 11,564 observations with 14 variables. This dataset includes detailed environmental conditions for each marathon, with runner performance represented by `Percent_CR`, indicating the percentage deviation from the current course record. Additional datasets include air quality index (AQI) values, which were obtained from the Environmental Protection Agency's (EPA) Air Quality System (AQS) using the `RAQSAPI` R package. The AQI data were collected as daily summaries for specific parameters, such as PM2.5 and ozone levels, corresponding to the Core-Based Statistical Areas (CBSAs) of the marathon locations on their respective race dates. Other datasets include records of course performances and annual marathon dates, which were integrated to provide a comprehensive view of historical race performance and environmental factors.

Exploratory Data Analysis

Data quality check is performed on all variables of the given dataset. The column `Flag`, which are weather indicators based on WBGT values, has `NA` values stored as empty strings. We inspected the frequencies of the column and noticed that there are in total 491 empty strings, which then we converted to `NA` values for further analysis. The variable `X.rh`, which is the percent relative humidity, should have values in percentage unit (e.g., 38). However, we noticed that a substantial amount of values are between 0 and 1, which is not correctly recorded as it is impossible to have a humidity near 0%. To assure accurate representation of humidity, a new column `Percent_RH` is mutated to represent humidity, where the inaccurate values were corrected by multiplying 100 to have consistent percentage format as the other values in the column.

The dataset contains 14 variables, in which 2 variable (`marathon_name`, `Year`) are information related to the marathon course, 2 variables (`Sex`, `Age`) are runners' characteristics, `Percent_CR`, which is the percent off current record indicates runners' performances, and the rest of the variables (`Flag`: indicator based on WBGT, `Td..C`: Dry bulb temperature in Celsius, `Tw..C`: Wet bulb temperature in Celsius, `Percent_RH`: Percent relative humidity, `Tg..C`: Black globe temperature in Celsius, `SR.W.m2`: Solar radiation in Watts per meter squared, `DP`: Dew Point in Celsius, `Wind`: Wind speed in Km/hr, `WBGT`: Wet Bulb Globe Temperature) are estimators of weather conditions.

Exploratory analysis is first carried out on personal characteristics. Table 1 shows there are no missing values in these columns. Age and %CR have similar distributions across different marathon race. Performances (percent off current record) is similar across the four races except for Boston. The obvious faster Percent CR for Boston is likely due to the fact that the Boston Marathon requires runners to meet a qualifying time based on their age and gender.

Table 1: Summary Table of Runner Characteristics

Characteristic	**Boston** **N = 2,088**	**Chicago** **N = 2,553**	**Grandmas** **N = 2,000**	**NYC** **N = 2,930**	**Twin Cities** **N = 1,993**
Sex					
Female	984 (47%)	1,210 (47%)	934 (47%)	1,402 (48%)	922 (46%)
Male	1,104 (53%)	1,343 (53%)	1,066 (53%)	1,528 (52%)	1,071 (54%)
Year	2,008 (2,003, 2,012)	2,006 (2,001, 2,011)	2,008 (2,004, 2,012)	2,004 (1,998, 2,010)	2,008 (2,004, 2,012)
Age	47 (32, 61)	46 (30, 61)	44 (29, 58)	49 (33, 65)	44 (30, 59)
Percent off current course record	32 (18, 56)	38 (20, 67)	38 (20, 62)	37 (19, 69)	36 (19, 63)

[1] n (%); Median (Q1, Q3)

Summary characteristics of weather conditions in each marathon is presented in Table 2. Variables `Td..C`,`Tw..C` and `Tg..C` are not included here as they are used in the calculation of `WBGT`. From the table, we can see that each marathon has data for approximately 20 years in the dataset. The distribution of `Flag`, which is bins calculated based on WBGT and risk of heat illness, is different for the 5 marathons. We can observe similar pattern in `WBGT`, where Grandmas has the highest average `WBGT` (18.1) and Boston has the lowest (9.9), indicating the potential need to stratify our analysis by marathon race. Missing values are also observed in marathons except Boston, which will be explored further in the next section.

Table 2: Summary Characteristics of Weather Parameters

Characteristic	**Boston** **N = 18**	**Chicago** **N = 21**	**Grandmas** **N = 17**	**NYC** **N = 23**	**Twin Cities** **N = 17**
Year	2,008 (2,003, 2,012)	2,006 (2,001, 2,011)	2,008 (2,004, 2,012)	2,004 (1,998, 2,010)	2,008 (2,004, 2,012)
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%)
Unknown	0	1	1	1	1
Solar radiation in Watts per meter squared	721 (574, 800)	470 (437, 518)	736 (571, 838)	393 (309, 546)	488 (355, 541)
Unknown	0	1	1	1	1
Percent relative humidity	37 (1, 58)	60 (53, 68)	58 (1, 78)	1 (0, 55)	53 (1, 70)
Unknown	0	1	1	1	1
Dew Point in Celsius	3 (0, 6)	6 (-2, 10)	12 (11, 14)	2 (-4, 9)	6 (3, 10)
Unknown	0	1	1	1	1
Wind speed in Km/hr	11.8 (8.3, 16.0)	8.0 (5.3, 10.2)	9.3 (7.7, 11.2)	11.2 (9.0, 14.0)	9.3 (6.5, 10.0)
Unknown	0	1	1	1	1
Wet Bulb Globe Temperature (WBGT)	9.9 (8.7, 12.7)	13.1 (7.2, 16.1)	18.1 (16.0, 21.0)	10.2 (6.7, 14.1)	12.6 (9.0, 16.3)
Unknown	0	1	1	1	1

[1] Median (Q1, Q3); n (%)

The summary table below provides a detailed comparison of AQI values across five cities. The parameter codes used to measure AQI are codes 44201 (Ozone), 88101 (PM2.5) and 88502 (PM2.5). The units of measurement primarily include micrograms per cubic meter (LC) and parts per million. Sample durations include 1-hour, 24-hour, and 8-hour averages. The ozone AQI is measured by 8-hour sample duration and particle pollution is measured by 24-hour sample duration.

Table 3: Summary Table of AQI Values

Characteristic	**Boston** **N = 1,693**	**Chicago** **N = 3,200**	**Grandmas** **N = 422**	**NYC** **N = 3,905**	**Twin Cities** **N = 1,231**
parameter_code					
44201	764 (45%)	1,657 (52%)	180 (43%)	1,687 (43%)	256 (21%)
88101	800 (47%)	1,370 (43%)	211 (50%)	1,764 (45%)	873 (71%)
88502	129 (7.6%)	173 (5.4%)	31 (7.3%)	454 (12%)	102 (8.3%)
units_of_measure					
Micrograms/cubic meter (LC)	929 (55%)	1,543 (48%)	242 (57%)	2,218 (57%)	975 (79%)
Parts per million	764 (45%)	1,657 (52%)	180 (43%)	1,687 (43%)	256 (21%)
sample_duration					
1 HOUR	275 (16%)	525 (16%)	69 (16%)	643 (16%)	162 (13%)
24 HOUR	594 (35%)	1,194 (37%)	117 (28%)	1,693 (43%)	436 (35%)
24-HR BLK AVG	251 (15%)	239 (7.5%)	101 (24%)	303 (7.8%)	441 (36%)
8-HR RUN AVG BEGIN HOUR	573 (34%)	1,242 (39%)	135 (32%)	1,266 (32%)	192 (16%)
aqi	41 (33, 48)	38 (29, 55)	32 (23, 42)	28 (23, 39)	26 (18, 38)

Table 3: Summary Table of AQI Values *(continued)*

Characteristic	**Boston** N = 1,693	**Chicago** N = 3,200	**Grandmas** N = 422	**NYC** N = 3,905	**Twin Cities** N = 1,231
Unknown	275	525	69	643	162
arithmetic_mean	1.8 (0.0, 7.1)	0.0 (0.0, 9.1)	2.9 (0.0, 6.9)	2.9 (0.0, 6.6)	3.8 (1.8, 6.6)

[1] n (%); Median (Q1, Q3)

Missing Data Pattern

The dataset contains missing values, with approximately 4% of all variables missing, evenly distributed across weather parameters. The missingness occurs in the same rows, as evidenced by the fact that 491 rows have `NA` values across all weather parameters, matching the number of missing observations for each parameter individually.

Analysis of the missing data reveals that these rows are concentrated in the years 2011 and 2012. This suggests that the missingness is not completely random but is instead related to specific years, likely due to external factors such as unavailability of weather records during that period. Consequently, this missingness is classified as Missing At Random (MAR), as the probability of data being missing is dependent on the observed variable `Year` but not on the missing weather parameters themselves.

Table 3 provides a summary of the characteristics of the missing and non-missing groups, showing that their distributions are similar except for the parameter `Year`.

Table 4: Missing Values Pattern

Characteristic	**Observations with Missing weather variables** **N = 491**	**Observations with non-missing weather variables** **N = 11,073**
marathon_name		
Boston	0 (0%)	2,088 (19%)
Chicago	126 (26%)	2,427 (22%)
Grandmas	116 (24%)	1,884 (17%)
NYC	131 (27%)	2,799 (25%)
Twin Cities	118 (24%)	1,875 (17%)
Sex		
Female	234 (48%)	5,218 (47%)
Male	257 (52%)	5,855 (53%)
Year		
2011	375 (76%)	241 (2.2%)
2012	116 (24%)	363 (3.3%)
Age	46 (31, 62)	46 (31, 61)
Percent_CR	37 (19, 61)	36 (19, 63)

[1] n (%); Median (Q1, Q3)

Effects of Increasing Age on Marathon Performance Across Gender

Figure 1 shows a U-shaped pattern in performance across all marathons. Runners reach their fastest times at younger ages and then gradually slow down as they get older. The decline in performance is steeper for females than for males as they age. Additionally, marathons in Chicago and NYC show the most noticeable slowing compared to the other races.

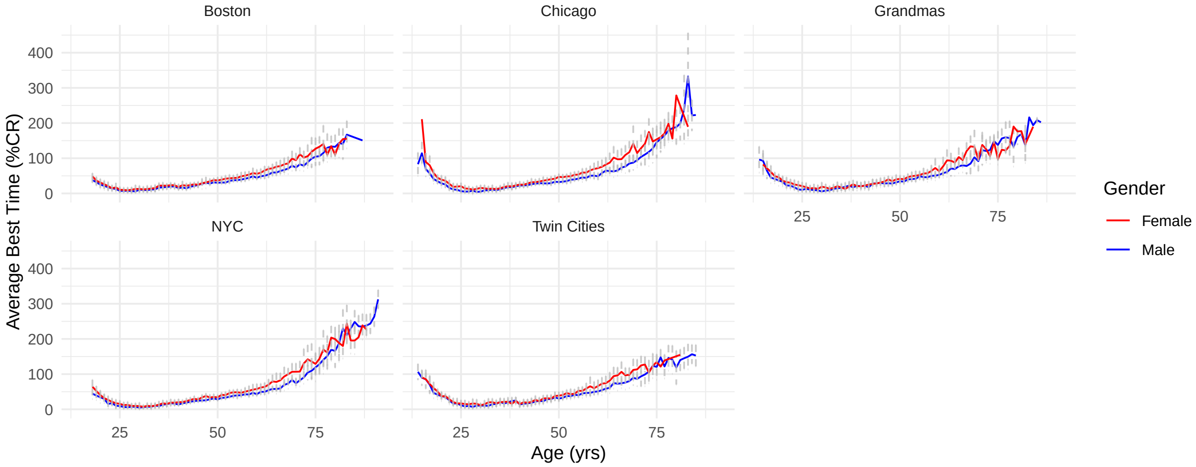

Figure 1: Effect of Age on Marathon Performances

Age of Peak Performances

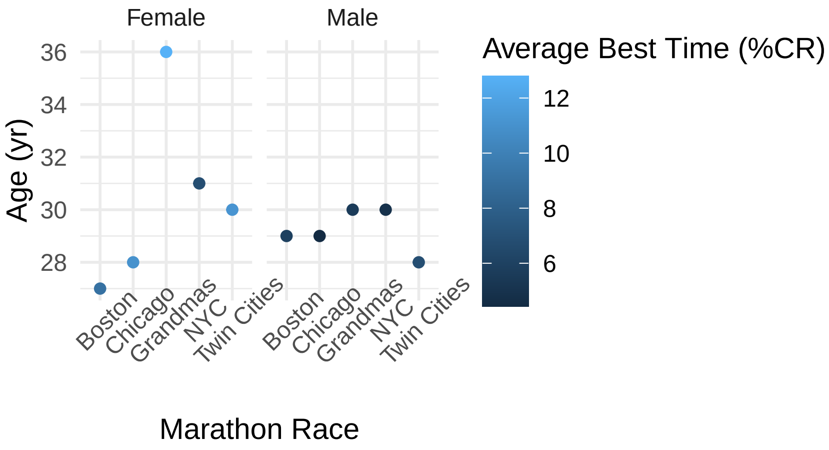

Figure 2: Fastest Running Age for Women and Men

Figure 2 shows that the fastest marathon running age is around 30 years old for both male and females across all five races. However, males show more consistency in their peak running ages across the marathons, with less variation compared to females. For females, the fastest running ages tend to vary more and are slightly higher than those of males. Overall, both genders share a similar range for their peak performance age, but males exhibit more stability across different marathon races.

Age of Significant Slowing

To identify the age where performance slowing becomes significant, we calculated the slope of percent off current record (%CR) change using a window size of five years. Based on the distribution of slopes, we set a threshold of 5, defining the age where the slope exceeds this threshold as the age of significant slowing. At this point, the rate of performance decline is five times the rate of age increase. From Table 4, the age of significant slowing is identified as 60 for women and 66 for men. This pattern is further illustrated in the accompanying plot, where the slopes (scaled by 5 for visualization purposes) are plotted alongside %CR change with age. The plot shows that after the identified ages, the performance slowing becomes more pronounced and unstable for both genders. Additionally, the slightly earlier slowing for women compared to men aligns with the observed %CR trend.

Table 5: Age of Significant Slowing For Women and Men

Age of significant slowing	Average %CR	Slope	Sex
66	70.17	5.40	Male
60	62.94	5.44	Female

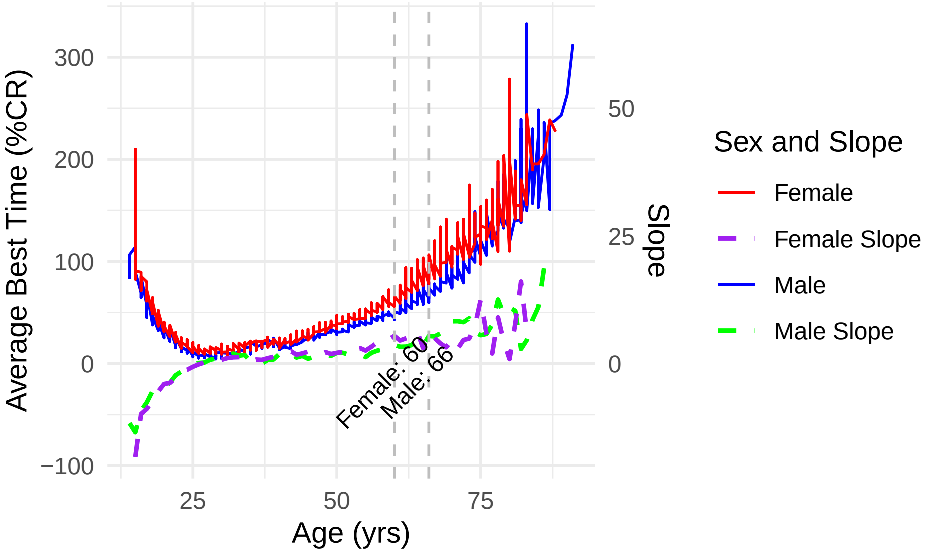

Figure 3: Significant Slowing Age for Women and Men: %CR and Slopes

Impact of environmental conditions on marathon performance

Air Quality Impact

The AQI estimates include durations of 1-hour, 24-hour, and 8-hour run averages. To reflect air quality conditions on race day, we calculated AQI values by averaging the 24-hour (PM2.5) and 8-hour (ozone) estimates. To account for differences in age, we stratified runners into four age categories: 14–31, 32–46, 47–61, and 62–91, and examined the effect of air quality on performance (%CR) for each group by sex.

The plot demonstrates that air quality impacts older runners (age group 62–91) more significantly than younger groups, as evident by the higher %CR values across AQI levels and the increasing %CR trend when AQI levels increase. This trend is consistent for both males and females. For younger runners (14–31 and 32–46), %CR remains relatively stable across AQI values, indicating that air quality has less of an impact on their performance. Additionally, mid-aged runners (47–61) show a moderate sensitivity to AQI changes, with a slight upward trend in %CR as AQI values increase.

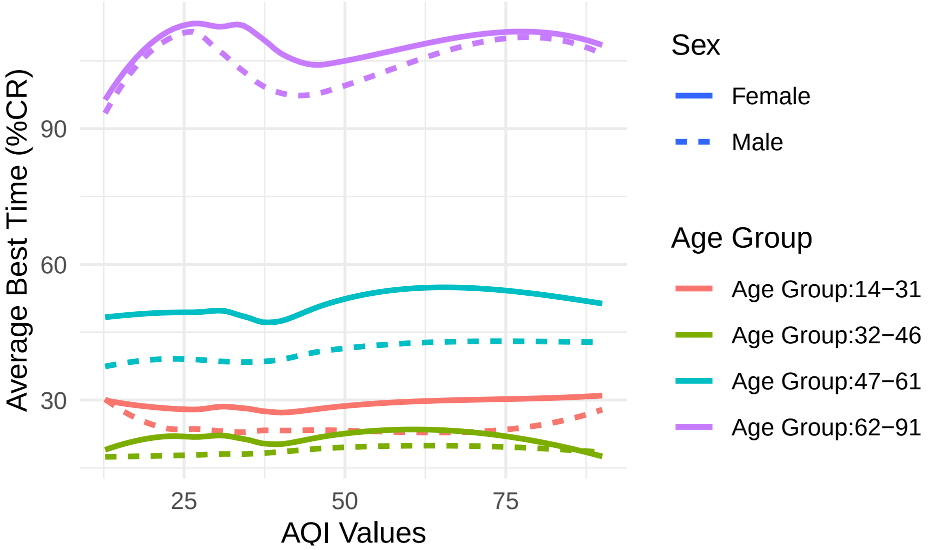

Figure 4: Effect of Air Quality on Marathon Performances

Weather Parameters Impact

There are in total 8 continuous weather parameters in `projct1` dataset. These vairables are `Td..C`, `Tw..C`, `Percent_RH`, `Tg..C`, `SR.W.m2`, `DP`, `Wind`, and `WBGT`. In the preliminary research, analyses were carried out on how `WBGT` impacts runners' performances. To have a more comprehensive understanding of the overall weather's impact on runner's performances, correlation between the weather variables are needed for further analyses.

The heatmap represents the coreelation between the weather variables. `WBGT` (Wet Bulb Globe Temperature) is strongly correlated with `Tw..C` (wet bulb temperature), `Td..C` (dry bulb temperature), and `Tg..C` (globe temperature). Because of this high correlation, focusing on `WBGT` makes sense as a single indicator of overall temperature estimates. However, `Percent_RH` (relative humidity), `SR.W.m2` (solar radiation) and `Wind` show low correlation with `WBGT` and other temperature variables, indicating that these factors provide independent information about weather conditions and should be considered separately as important factors when analyzing their impact on performance.

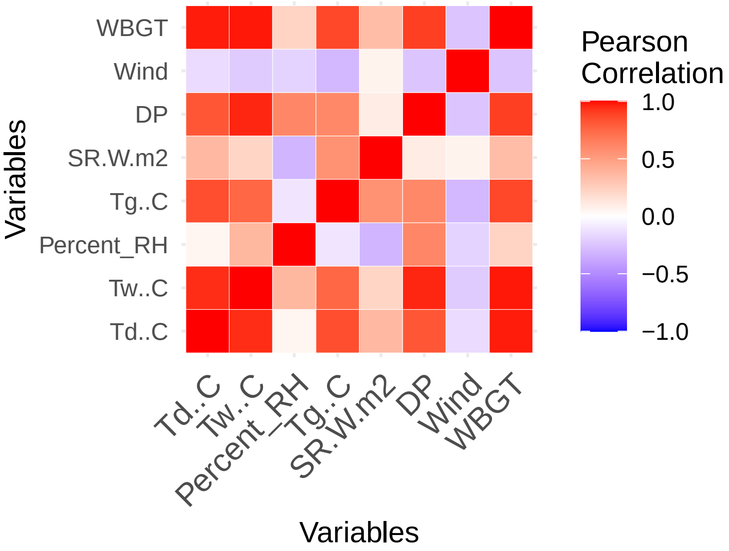

Figure 5: Correlation Heatmap

The following plots illustrate the effect of various weather parameters—Wet Bulb Globe Temperature (WBGT), humidity, solar radiation, and wind—on marathon performance (%CR) across different age groups and by sex. WBGT has a pronounced impact, particularly on older runners (age group 62–91), with %CR increasing more steeply as WBGT rises above 10, indicating greater performance slowing in this group. This trend is more pronounced in females than males. For humidity, the performance of older runners also shows a stronger sensitivity, while younger groups (14–31 and 32–46) exhibit relatively stable performance. Similarly, solar radiation disproportionately affects older runners, with %CR showing an unstable fluctuation as solar radiation increases, whereas younger runners remain less affected, with slight upward trend in %CR as solar radiation increases. Lastly, wind has a moderate effect, with %CR showing a slight increase for older groups as wind speed rises, although the impact is less distinct compared to WBGT and solar radiation. Overall, the results consistently show that older runners, particularly females, are more sensitive to adverse weather conditions, while younger age groups maintain better performance stability across varying environmental factors.

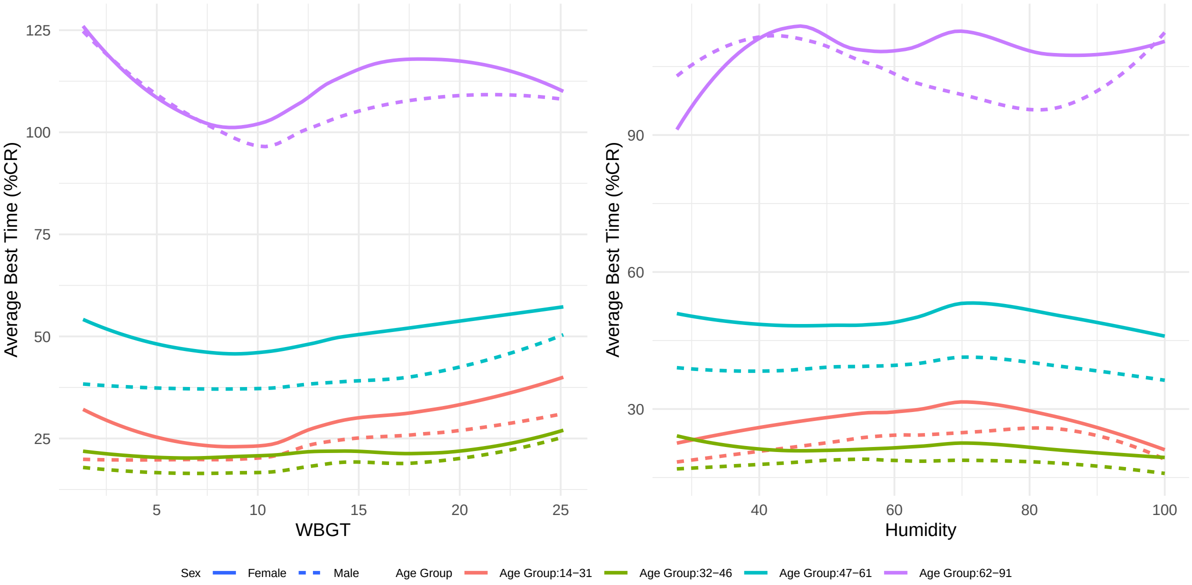

Figure 6: Effect of WBGT and Humidity on Marathon Performance

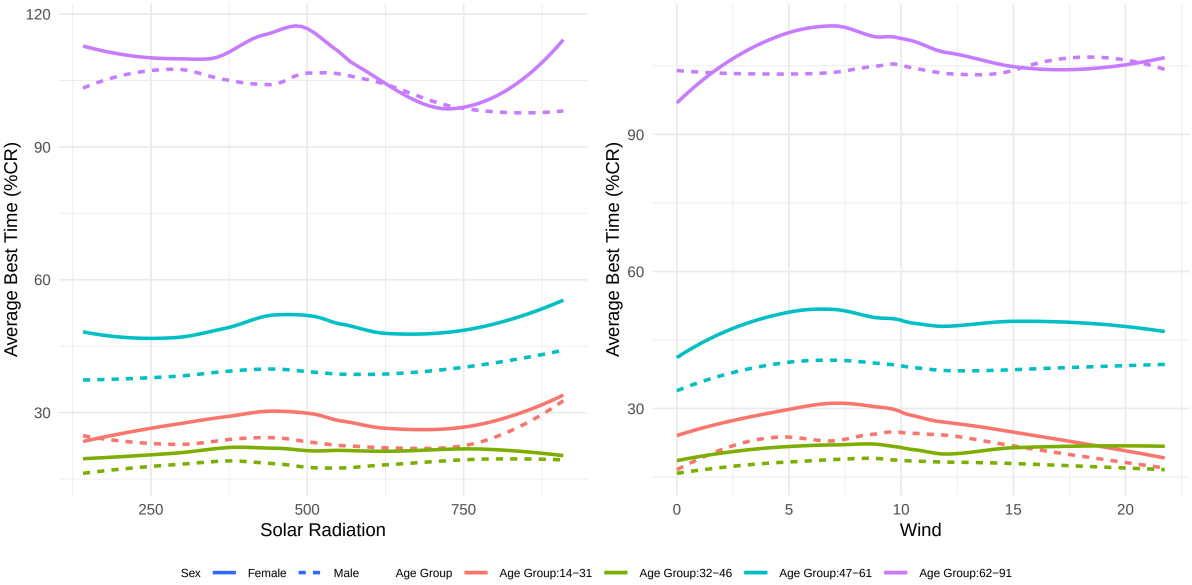

Figure 7: Effect of Solar Radiation and Wind on Marathon Performance

Regression Analysis

Given the exploratory data analysis (EDA) plots, we observed significant weather impacts, particularly on the performance of older age groups, as well as noticeable differences in performance trends between female and male runners. To investigate these observations statistically, we fitted two separate linear models: one for female runners and one for male runners. The model structure allows us to assess the effects of various weather parameters, age, and their interactions on performance (%CR) for each sex.

$$\begin{aligned}\text{Percent\_CR} = {} & \beta_0 + \beta_1 \cdot \text{Percent\_RH} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Age}^2 \\ & + \beta_4 \cdot \text{SR.W.m2} + \beta_5 \cdot \text{Wind} + \beta_6 \cdot \text{WBGT} + \beta_7 \cdot \text{aqi\_values} \\ & + \beta_8 \cdot \text{Percent\_RH} \cdot \text{Age} + \beta_9 \cdot \text{SR.W.m2} \cdot \text{Age} + \beta_{10} \cdot \text{Wind} \cdot \text{Age} \\ & + \beta_{11} \cdot \text{WBGT} \cdot \text{Age} + \beta_{12} \cdot \text{aqi\_values} \cdot \text{Age} + \epsilon\end{aligned}$$

From the model for male runners, significant predictors included humidity(`Percent_RH`), age (`age`), quadratic form of age (`Age^2`), solar radiation (`SR.W.m2`), interaction term between humidity and age (`Percent_RH:Age`), interaction term between solar radiation and age (`Age:SR.W.m2`), and interaction term between WBGT and age (`Age:WBGT`). These results indicate that age and its quadratic term have the strongest effects on performance, with interactions between weather parameters and age further influencing outcomes. For female runners, significant predictors included age (`Age`), quadratic form of age (`Age^2`), and WBGT (`WBGT`), highlighting the dominant role of age and WBGT on performance for women. Notably, the quadratic term (`Age^2`) is significant in both models, confirming the nonlinear relationship between age and performance observed in the EDA plots. However, fewer interaction terms were significant in the female model, suggesting different sensitivities to environmental conditions compared to male runners.

These results align with the trends observed in the EDA plots, confirming that age and environmental conditions, particularly WBGT, significantly affect performance, with nuanced differences between sexes.

Table 6: Significant Variables in Male Runners Model

	Estimate	**Std. Error**	**t value**	**P Value**
(Intercept)	104.2046068	5.2272645	19.934826	0.0000000
Percent_RH	0.1541018	0.0498661	3.090310	0.0020089
Age	-5.6102979	0.1261042	-44.489398	0.0000000
I(Age^2)	0.0805839	0.0007974	101.062154	0.0000000
SR.W.m2	0.0096378	0.0046016	2.094440	0.0362638
Percent_RH:Age	-0.0046325	0.0009671	-4.790054	0.0000017
Age:SR.W.m2	-0.0002930	0.0000909	-3.225075	0.0012663
Age:WBGT	0.0085365	0.0031177	2.738109	0.0061981

Table 7: Significant Variables in Female Runners Model

	Estimate	**Std. Error**	**t value**	**P Value**
(Intercept)	125.0744802	5.4759474	22.840702	0.0000000
Age	-5.9026067	0.1423313	-41.470895	0.0000000
I(Age^2)	0.0817898	0.0009476	86.309347	0.0000000
WBGT	0.3422804	0.1647502	2.077572	0.0377977

Discussion

This report utilizes multiple exploratory data analysis (EDA) methods, including missing data detection, summary tables, visualizations, and linear regression modeling, to investigate the relationship between weather conditions and marathon performances, with a focus on variations across age and gender. The findings reveal that age is the most significant factor influencing marathon performance. Both males and females exhibit similar trends, with performance peaking around age 30 and slowing down significantly after approximately 60 years old. However, the age effect is slightly more pronounced in females, with significant slowing observed at 60 years old compared to 66 years old for males.

Regarding air quality, its impact is predominantly observed in older age groups, while younger runners' performances remain stable under varying air quality levels. The differences between sexes in response to air quality were minimal. Weather parameters, particularly WBGT, humidity, and solar radiation, exhibited the strongest influence on performance. These effects were most pronounced in older runners, where %CR (indicating worse performance) fluctuates and show an increasing trend as these weather parameters rose. Conversely, younger runners displayed relatively stable %CR values, showing limited sensitivity to changes in weather conditions. Statistical models further supported these findings, indicating differences in sensitivity to weather parameters between males and females. However, WBGT and age were significant predictors in both models, underscoring their critical role in affecting performance.

One limitation of our study is we did not analyze the same cohort of runners across all years and marathons, which makes it challenging to control for various confounding factors such as individual runner performance, training, and environmental adaptations. Additionally, our sample size was smaller for older age groups, potentially affecting the reliability of results for those ages. Furthermore, as our objective was primarily exploratory data analysis (EDA), marginal differences in performance may be difficult to distinguish using plots and summary tables alone. More thorough analyses, such as more sophisticated regression model or advanced machine learning models, may be necessary to obtain more precise and accurate results.

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.

Appendix

```r
# lIbrary used in this analysis
library(dplyr)
library(knitr)
library(kableExtra)
library(visdat)
library(ggplot2)
library(gtsummary)
library(reshape2)
library(car)
library(RColorBrewer)
library(ggpubr)
# Read in datasets
project1 <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Desktop/PHP2550/Proje
aqi_values <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Desktop/PHP2550/Pro
course_record <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Desktop/PHP2550/
marathon_dates <- read.csv("~/Library/Mobile Documents/com~apple~CloudDocs/Desktop/PHP2550
# Data pre-proccessing

# Add a new column two columns (marathon_name and Sex) for EDA purpose
project1  <- project1 %>%
      mutate(marathon_name = case_when(Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.==0 ~
                                         "Boston",
                                   Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.==1~
                                     "Chicago",
                                   Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.==2 ~
                                     "NYC",
                                   Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.==3 ~
                                     "Twin Cities",
                                   TRUE ~ "Grandmas")) %>%
  mutate(Sex=case_when(`Sex..0.F..1.M.`==1~"Male", TRUE ~ "Female"))

# Factor Sex column
project1$Sex <- as.factor(project1$Sex)

# Change the name of columns 'Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.',
# 'Age..yrs', and 'X.CR'
names(project1)[names(project1) %in% c("Age..yr.", "X.CR")] <-
  c("Age", "Percent_CR")

# Delete Sex..0.F..1.M and `Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.` column
project1 <- project1 %>%
  dplyr::select(-`Sex..0.F..1.M.`,
                -`Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.` )

# Covert empty strings to NA values for Flag column
project1["Flag"][project1["Flag"] == ""] <- NA

# Covert incorrect values in humidify column
project1<- project1 %>%
  mutate(Percent_RH = ifelse(X.rh <= 1, X.rh * 100, X.rh))
# Create a data frame with only columns related to runner characteristics
runner_data <- project1 %>%
  dplyr::select(marathon_name,
         Sex,
         Year,
         Age,
         Percent_CR)
# Create a summary table of the runner characteristics
tbl_summary_runner <- runner_data %>%
  tbl_summary(
    by = marathon_name,
    label = c(Sex ~ "Sex",
              Year ~ "Year",
              Age ~ "Age",
              Percent_CR ~ "Percent off current course record"),
    statistic = list(all_continuous() ~ "{median} ({p25}, {p75})",
                     all_categorical() ~
    "{n} ({p}%)"),
    missing = "ifany") %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Table of Runner Characteristics",
longtable = TRUE, linesep = "") %>%
kableExtra::kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
column_spec(1, width = "4cm") %>%
   column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE)

tbl_summary_runner



# Create a dataframe with only columns related to weather
weather_data_summary <- project1 %>%
  dplyr::select(-Sex,-Age,-Percent_CR) %>%

  # Group by marathon race and year
  group_by(marathon_name, Year) %>%

  # Calculate mean values of weather parameters
  summarise(Flag=Flag, SR.W.m2=mean(SR.W.m2), X.rh=mean(X.rh), DP=mean(DP),
            Wind=mean(Wind), WBGT=mean(WBGT)) %>%
  unique()

# Create a summary table of the weather variables
tbl_summary_weather <- weather_data_summary %>%
  tbl_summary(
    by = marathon_name,
    label = c(Year ~ "Year",
              Flag ~ "Flag",
              SR.W.m2 ~ "Solar radiation in Watts per meter squared",
              X.rh ~ "Percent relative humidity",
              DP ~ "Dew Point in Celsius",
              Wind ~ "Wind speed in Km/hr",
              WBGT ~ "Wet Bulb Globe Temperature (WBGT)"),
    statistic = list(all_continuous() ~ "{median} ({p25}, {p75})",
                     all_categorical() ~
    "{n} ({p}%)"),
    missing = "ifany") %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Characteristics of Weather Parameters
longtable = TRUE, linesep = "") %>%
kableExtra::kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
column_spec(1, width = "4cm") %>%
   column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE)
tbl_summary_weather

# Create a summary table of the AQI values
tbl_summary_aqi <- aqi_values %>% select(parameter_code, units_of_measure, sample_duration
                                         arithmetic_mean, marathon) %>%
  tbl_summary(
    by = marathon,
    statistic = list(all_continuous() ~ "{median} ({p25}, {p75})",
                     all_categorical() ~
    "{n} ({p}%)"),
    missing = "ifany") %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Table of AQI Values",
longtable = TRUE, linesep = "") %>%
kableExtra::kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down"))  %>%
column_spec(1, width = "4cm") %>%
   column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE)
```

```
tbl_summary_aqi
# Mutate a column to indicate missingness status
project1_with_missing_status <- project1 %>%
  mutate(missingness=case_when(if_all(c("Td..C","Tw..C","X.rh","Tg..C",
                                        "SR.W.m2","DP","Wind","WBGT"),
                                      is.na) ~
                                 "Observations with Missing weather variables",
                               TRUE ~ "Observations with non-missing weather variables"))

# Factor year column
project1_with_missing_status$Year <- as.factor(project1_with_missing_status$Year)

# Create a summary table comparing variables between variables of missing group and
# variables of non-missing group
tbl_summary_missing <- project1_with_missing_status[, c("marathon_name", "Sex", "Year",
                                                        "Age", "Percent_CR",
                                                        "missingness")] %>%
  tbl_summary(
    by = missingness,
    statistic = list(all_continuous() ~ "{median} ({p25}, {p75})",
                     all_categorical() ~
    "{n} ({p}%)"),
    missing = "no") %>%
  modify_table_body(
    # Retain only rows with Year as 2011 or 2012
    ~ .x %>%
      filter(!label %in% 1993:2010) %>%
      filter(!label %in% 2013:2016)
  ) %>%
  as_kable_extra(booktabs = TRUE, caption = "Missing Values Pattern",
longtable = TRUE, linesep = "") %>%
kableExtra::kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
column_spec(1, width = "3cm") %>%
   column_spec(2, width = "5cm") %>%
  column_spec(3, width = "5cm") %>%
  row_spec(0, bold = TRUE)

tbl_summary_missing
# Create age dataframes for women and men ready for plotting
age_data_male <- project1 %>%
  filter(Sex=="Male") %>%
  group_by(marathon_name, Age) %>%
  summarise(mean_CR = mean(Percent_CR, na.rm=T), se_CR = sd(Percent_CR, na.rm=T))
age_data_female <- project1 %>%
  filter(Sex=="Female") %>%
  group_by(marathon_name, Age) %>%
  summarise(mean_CR = mean(Percent_CR, na.rm=T), se_CR = sd(Percent_CR, na.rm=T))

# Create a %CR vs. Age plot stratified by marathon race and gender
age_plot <- ggplot() +
  geom_line(data=age_data_male, aes(x = Age, y = mean_CR, color="Male"), size=0.5) +
  geom_errorbar(data=age_data_male, aes(x=Age, ymin = mean_CR -se_CR,
                    ymax = mean_CR +se_CR),
                color = "grey", width = 0.1, linetype="dashed", alpha=0.8) +
  geom_line(data=age_data_female, aes(x = Age, y = mean_CR, color="Female"), size=0.5) +
  geom_errorbar(data=age_data_female, aes(x=Age,ymin = mean_CR -se_CR,
                    ymax = mean_CR +se_CR),
                color = "grey", width = 0.1, linetype="dashed", alpha=0.8) +
  facet_wrap(~marathon_name)+
  theme_minimal()+
  scale_color_manual(values = c("Male" = "blue", "Female" = "red"))+
  labs(color="Gender",
       x="Age (yrs)", y="Average Best Time (%CR)")

age_plot


# Get the age of fastest running for male
fastest_male <- age_data_male %>%
  group_by(marathon_name) %>%
  filter(mean_CR== min(mean_CR, na.rm = TRUE)) %>%
  dplyr::select(marathon_name, Age, mean_CR)

# Get the age of fastest running for female
fastest_female <- age_data_female %>%
  group_by(marathon_name) %>%
  filter(mean_CR== min(mean_CR, na.rm = TRUE)) %>%
  dplyr::select(marathon_name, Age, mean_CR)

# Combine the age for both men and women into one dataset for plotting
fastest_table <- rbind(fastest_male, fastest_female)
```

```r
fastest_table$Sex <- rep(c("Male","Female"), each=5)
# Visualize fastest running age for men and women
fastest_plot <- ggplot(data=fastest_table) +
  geom_point(aes(x=marathon_name, y=Age, color=mean_CR))+
  facet_wrap(~Sex)+
  theme_minimal()+
  theme(axis.text.x = element_text(angle=45))+
  labs(x="Marathon Race",
       y="Age (yr)",
       color = "Average Best Time (%CR)")
fastest_plot
# Define window size for slope calculation (5 years)
window_size <- 5

# Create a function to calculate slopes over a moving window
calculate_slope <- function(data) {
  data %>%
    arrange(Age) %>%
    mutate(
      slope = (lead(avg_CR, window_size) - avg_CR) / window_size
    )
}

# Calculate mean %CR for both men and women across 5 races
age_data_male_by_age <- age_data_male %>%
  group_by(Age) %>%
  summarise(avg_CR = mean(mean_CR))
age_data_female_by_age <- age_data_female %>%
  group_by(Age) %>%
  summarise(avg_CR = mean(mean_CR))

# Calculate slopes for males and females
age_data_male_slopes <- calculate_slope(age_data_male_by_age)
age_data_female_slopes <- calculate_slope(age_data_female_by_age)

# Define threshold for significant slowing (e.g., slope > 5)
threshold <- 5

# Find the first age where slope exceeds the threshold
slowing_point_male <- age_data_male_slopes %>%
  filter(slope > threshold) %>%
  slice(1)

slowing_point_female <- age_data_female_slopes %>%
  filter(slope > threshold) %>%
  slice(1)

# Combine the two datasets
slowing_point_combined <- rbind(slowing_point_male, slowing_point_female)

# Add a new column to indicate Sex
slowing_point_combined$Sex <- c("Male", "Female")

# Round numbers
slowing_point_combined$slope <- round(slowing_point_combined$slope,2)
slowing_point_combined$avg_CR <-
round(slowing_point_combined$avg_CR,2)

# Change column names
colnames(slowing_point_combined) <- c("Age of significant slowing", "Average %CR",
                                      "Slope", "Sex")

slowing_point_combined <- slowing_point_combined %>%
  kable(booktabs = TRUE, caption = "Age of Significant Slowing For Women and Men",
        longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down"))

slowing_point_combined

# Create a %CR vs. Age plot stratified, add slopes
age_slope_plot <- ggplot() +
  geom_line(data=age_data_male, aes(x = Age, y = mean_CR, color="Male"), size=0.5) +
  geom_line(data=age_data_female, aes(x = Age, y = mean_CR, color="Female"), size=0.5) +
  geom_line(data = age_data_male_slopes, aes(x=Age, y = slope * 5, color = "Male Slope"),
            size = 0.8, linetype = "dashed") + # Green for male slope
  geom_line(data = age_data_female_slopes, aes(x=Age, y = slope * 5, color = "Female Slope
            size = 0.8, linetype = "dashed") + # Purple for female slope
  theme_minimal() +
  scale_color_manual(values = c("Male" = "blue", "Female" = "red",
                                "Male Slope" = "green", "Female Slope" = "purple")) +
  scale_y_continuous(
    name = "Average Best Time (%CR)",
    sec.axis = sec_axis(~ . / 5, name = "Slope") # Secondary y-axis for slopes
  ) +
  geom_vline(xintercept = slowing_point_male$Age, linetype = "dashed", color = "grey", siz
  geom_vline(xintercept = slowing_point_female$Age, linetype = "dashed", color = "grey", s
  # Annotate text below the vertical lines
  annotate("text", x = slowing_point_male$Age, y = -30, label = paste("Male:", slowing_poi
           color = "black", angle = 45, vjust = -0.5, size = 3) +
  annotate("text", x = slowing_point_female$Age, y = -30, label = paste("Female:", slowing
           color = "black", angle =45, vjust = -0.5, size = 3)+
  labs(color="Sex and Slope",
       x="Age (yrs)") +
  theme(axis.title.y.right = element_text(color = "black")) # Different color for secondar

age_slope_plot

# Calculate average aqi values for each year in each marathon
aqi_average <- aqi_values %>%
  filter(sample_duration %in% c("8-HR RUN AVG BEGIN HOUR", "24 HOUR", "24-HR BLK")) %>%
  group_by(marathon, date_local) %>%
  summarise(mean_aqi = mean(aqi, na.rm=TRUE))

# Extract the year from `data_local`
aqi_average <- aqi_average %>%
  mutate(Year = as.integer(format(as.Date(date_local, format = "%Y-%m-%d"), "%Y")))


# Join `project1` with `aqi_acerage` based on `marathon` and `Year`
project1 <- project1 %>%
  left_join(aqi_average %>% dplyr::select(-date_local),
            by = c("marathon_name" = "marathon", "Year" = "Year"))

# Add age group to the dataset
project1_with_age_group <- project1 %>%
  mutate(age_group = case_when(Age >= 14 & Age <= 31 ~ "Age Group:14-31",
                   Age >= 32 & Age <= 46 ~ "Age Group:32-46",
                   Age >= 47 & Age <= 61 ~ "Age Group:47-61",
                   Age >= 62 & Age <= 91 ~ "Age Group:62-91"))
# Create a data frame summarising %CR by air quality, sex and age
air_quality_data <- project1_with_age_group %>%
  group_by(age_group) %>%
  summarise(performance = mean(Percent_CR, na.rm = TRUE))

# VIsualize air quality across age and gender
air_quality_plot <- ggplot(data=project1_with_age_group)+
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Male", ],
              aes(x=mean_aqi, y=Percent_CR, color = age_group, linetype = "Male"), method
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Female", ],
              aes(x=mean_aqi, y=Percent_CR, color = age_group, linetype = "Female"), metho
  theme_minimal()+
  labs(color="Age Group",
       linetype = "Sex",
       x="AQI Values", y="Average Best Time (%CR)")
air_quality_plot
# Select weather columns
correlation_data <- project1 %>%
  dplyr::select(Td..C, Tw..C, Percent_RH, Tg..C, SR.W.m2, DP, Wind, WBGT)

# Calculate the correlation matrix
cor_matrix <- cor(correlation_data, use = "complete.obs")

# Melt the correlation matrix
cor_melt <- melt(cor_matrix)

# Create the heatmap
cor_heatmap <- ggplot(cor_melt, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  coord_fixed() +
  labs(x = "Variables", y = "Variables")

cor_heatmap

# Plot for WBGT impact on performances
WBGT_plot <- ggplot(data=project1_with_age_group)+
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Male", ],
              aes(x=WBGT, y=Percent_CR, color = age_group, linetype = "Male"), method = 'l
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Female", ],
              aes(x=WBGT, y=Percent_CR, color = age_group, linetype = "Female"),method = '
  theme_minimal()+
  theme(legend.text=element_text(size = 7),
        legend.title=element_text(size = 7)) +
  labs(color="Age Group",
       linetype = "Sex",
       x="WBGT", y="Average Best Time (%CR)")

# Plot for WBGT impact on performances
humidity_plot <- ggplot(data=project1_with_age_group)+
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Male", ],
              aes(x=Percent_RH, y=Percent_CR, color = age_group, linetype = "Male"), metho
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Female", ],
              aes(x=Percent_RH, y=Percent_CR, color = age_group, linetype = "Female"),meth
  theme_minimal()+
  theme(legend.text=element_text(size = 7),
        legend.title=element_text(size = 7)) +
  labs(color="Age Group",
       linetype = "Sex",
       x="Humidity", y="Average Best Time (%CR)")

# Plot for Solar Radiation impact on performances
sr_plot <- ggplot(data=project1_with_age_group)+
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Male", ],
              aes(x=SR.W.m2, y=Percent_CR, color = age_group, linetype = "Male"), method =
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Female", ],
              aes(x=SR.W.m2, y=Percent_CR, color = age_group, linetype = "Female"),method
  theme_minimal()+
  theme(legend.text=element_text(size = 7),
        legend.title=element_text(size = 7)) +
  labs(color="Age Group",
       linetype = "Sex",
       x="Solar Radiation", y="Average Best Time (%CR)")


# Plot for Wind impact on performances
wind_plot <- ggplot(data=project1_with_age_group)+
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Male", ],
              aes(x=Wind, y=Percent_CR, color = age_group, linetype = "Male"),
              method = 'loess', se = FALSE)+
```

```r
  geom_smooth(data=project1_with_age_group[project1_with_age_group$Sex == "Female", ],
              aes(x=Wind, y=Percent_CR, color = age_group, linetype = "Female"),
              method = 'loess',  se = FALSE)+
  theme_minimal()+
  theme(legend.text=element_text(size = 7),
        legend.title=element_text(size = 7)) +
  labs(color="Age Group",
       linetype = "Sex",
       x="Wind", y="Average Best Time (%CR)")

weather_plot1_combined <- ggarrange(WBGT_plot, humidity_plot,
                                   ncol = 2, nrow = 1,
                                 common.legend = TRUE, legend = "bottom")
weather_plot1_combined
weather_plot2_combined <- ggarrange(sr_plot, wind_plot,
                                   ncol = 2, nrow = 1,
                                 common.legend = TRUE, legend = "bottom")
weather_plot2_combined
# Stratify by sex
male_data <- project1[project1$Sex == "Male", ]
female_data <- project1[project1$Sex == "Female", ]

# Male model with interaction terms between weather parameters and age
male_model <- lm(Percent_CR ~ Percent_RH + Age + I(Age^2) + SR.W.m2 + Wind +
                   WBGT + mean_aqi +
                   Percent_RH:Age + SR.W.m2:Age + Wind:Age +
                   WBGT:Age + mean_aqi:Age,
                 data = male_data)

# Female model with interaction terms between weather parameters and age
female_model <- lm(Percent_CR ~ Percent_RH + Age + I(Age^2) + SR.W.m2 + Wind +
                     WBGT + mean_aqi +
                     Percent_RH:Age + SR.W.m2:Age + Wind:Age +
                     WBGT:Age + mean_aqi:Age,
                   data = female_data)

# Collect model summary
model_summary_male <- summary(male_model)
model_summary_female <- summary(female_model)

# Extract coefficients as a data frame
```

```
coef_df_male <- as.data.frame(model_summary_male$coefficients)
coef_df_female <- as.data.frame(model_summary_female$coefficients)

# Rename columns for readability
colnames(coef_df_male) <- c("Estimate", "Std. Error", "t value", "P Value")
colnames(coef_df_female) <- c("Estimate", "Std. Error", "t value", "P Value")

# Filter only significant variables (e.g., p < 0.05)
significant_vars_male <- coef_df_male[coef_df_male$`P Value` < 0.05, ]
significant_vars_female <- coef_df_female[coef_df_female$`P Value` < 0.05, ]

# Generate tables with kable
significant_vars_male %>%
  kable(booktabs = TRUE, caption = "Significant Variables in Male Runners Model", longtabl
        linesep = "") %>%
  kable_styling(font_size = 8,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))%>%
column_spec(1, width = "3cm") %>%
   column_spec(2, width = "3cm") %>%
  column_spec(3, width = "3cm") %>%
  column_spec(4, width = "3cm") %>%
  column_spec(5, width = "3cm") %>%
  row_spec(0, bold = TRUE)

significant_vars_female %>%
  kable(booktabs = TRUE, caption = "Significant Variables in Female Runners Model", longta
        linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))%>%
column_spec(1, width = "3cm") %>%
   column_spec(2, width = "3cm") %>%
  column_spec(3, width = "3cm") %>%
  column_spec(4, width = "3cm") %>%
  column_spec(5, width = "3cm") %>%
  row_spec(0, bold = TRUE)
```