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. The results show that both male and female runners' performances peak around 30 years old and the slowing becomes significant around 40 years old. 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

The weather conditions in a marathon can largely affect runner's performances, and this impact varies across gender and age. Past studies show that an increasing WBGT (Wet Bulb Global Temperature) results in slowing of marathon performances and the impact is more evident in male runners than in female runners. Variables such as air quality, WBGT, solar radiation, wind and humidity are important factors to estimate weather conditions. This report utilizes a data consisting of weather and runner information from 5 marathon races (Boston, Chicago, NYC, Grandmas, Twin Cities) across 15-20 years. Analyzing methods include data quality checks, missing data analysis, summary tables and plots as well as correlation inspection and a linear model.

Data Collection

There are in total 4 datasets available for this report. The main dataset used in this report contains top single-age performances from five major marathons (Boston, Chicago, NYC, Grandmas, Twin Cities) cross 15-20 years from age 14-85 in men and women, with detailed environmental conditions for each marathon. The other 3 datasets available are: a dataset containing information about air quality records; a dataset of course records and a dataset with each year's marathon race dates. In the main dataset, the performance of the runners are represented by `Percent_CR`, which means the percent off current course record.

Data Preprocessing

The main dataset contains 11564 observations and 14 variables. For preprocessing, a new column `marathon_name` is created to replace the original `Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.` which uses 0,1,2,3,4 to indicate marathon race. The new column `marathon_name` is coded as "Boston", "Chicago", "NYC", "Grandmas", and "Twin Cities". Another column `Sex` is created to replace the original `Sex..0.F..1.M.` column that uses 0 and 1 to indicate sex. The new column `Sex` is coded as "Male" and "Female", which is also converted to factor for future analysis. To avoid long and unclear column names, names of columns `Age..yr.` and `X.CR` are changed to `Age` and `Percent_CR`.

Exploratory Data Analysis

After data-preprocessing, 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, 3 variables (`Sex`, `Age`, `Percent_CR`) are runners' characteristics, 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_CR) 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_CR	32 (18, 56)	38 (20, 67)	38 (20, 62)	37 (19, 69)	36 (19, 63)
[1] n (%); Median (IQR)					

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 (1,999, 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
SR.W.m2	721 (576, 799)	470 (439, 509)	736 (597, 835)	393 (309, 545)	488 (359, 539)
Unknown	0	1	1	1	1
X.rh	37 (1, 57)	60 (53, 67)	58 (1, 76)	1 (0, 55)	53 (1, 67)
Unknown	0	1	1	1	1
DP	3 (1, 6)	6 (0, 10)	12 (11, 14)	2 (-3, 8)	6 (3, 10)
Unknown	0	1	1	1	1

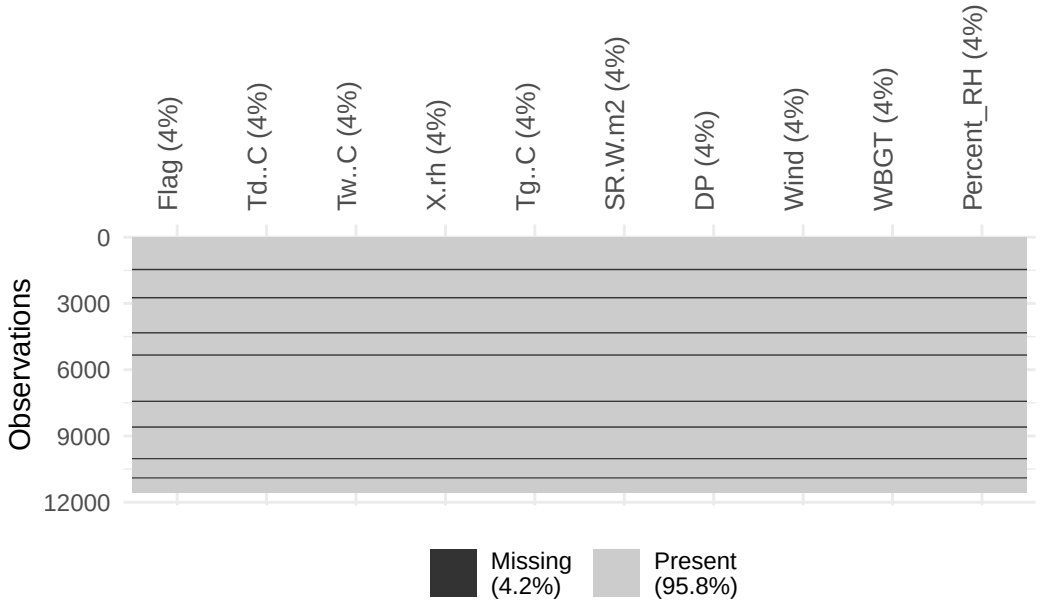
Table 2: Summary Characteristics of Weather Parameters *(continued)*

Characteristic	**Boston, N = 18**	**Chicago, N = 21**	**Grandmas, N = 17**	**NYC, N = 23**	**Twin Cities, N = 17**
Wind	11.8 (8.8, 15.4)	8.0 (5.3, 10.1)	9.3 (8.0, 11.2)	11.2 (9.2, 13.8)	9.3 (6.6, 10.0)
Unknown	0	1	1	1	1
WBGT	9.9 (8.8, 12.6)	13.1 (7.4, 16.0)	18.1 (16.6, 20.6)	10.2 (6.8, 13.9)	12.6 (9.1, 15.4)
Unknown	0	1	1	1	1

[1] Median (IQR); n (%)

Missing Data Pattern

Below is a visualization of the missingness in the dataset. The visualization indicates that all variables have a similar proportion of missing values (4%), suggesting that the missing data is evenly distributed across these variables. To validate that missingness occurs in the same rows, the number of rows (491) with all `NA` values in weather parameters ("Td..C","Tw..C","X.rh","Tg..C","SR.W.m2","DP","Wind","WBGT","Flag") is found to be equal to the number of missingness in each parameter (491), suggesting that missingness of the weather parameters occur in the same group of observations.



To identify what type of missingness this is, characteristics of the missing data and non-missing data are shown in Table 3. Characteristics of the two groups are similar except for `Year`. In the observations with missing weather variables, all missing rows are concentrated in 2011 and 2012, indicating that the reason for missing data is related to those years specifically, rather than being completely random. This is a type of Missing At Random (MAR) as the probability

of data being missing is related to observed data (in this case, the year), but not the missing values themselves.

Table 3: 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 (IQR)		

Effects of Increasing Age on Marathon Performance Across Gender

Figure 1: Effect of Age on Marathon Performances

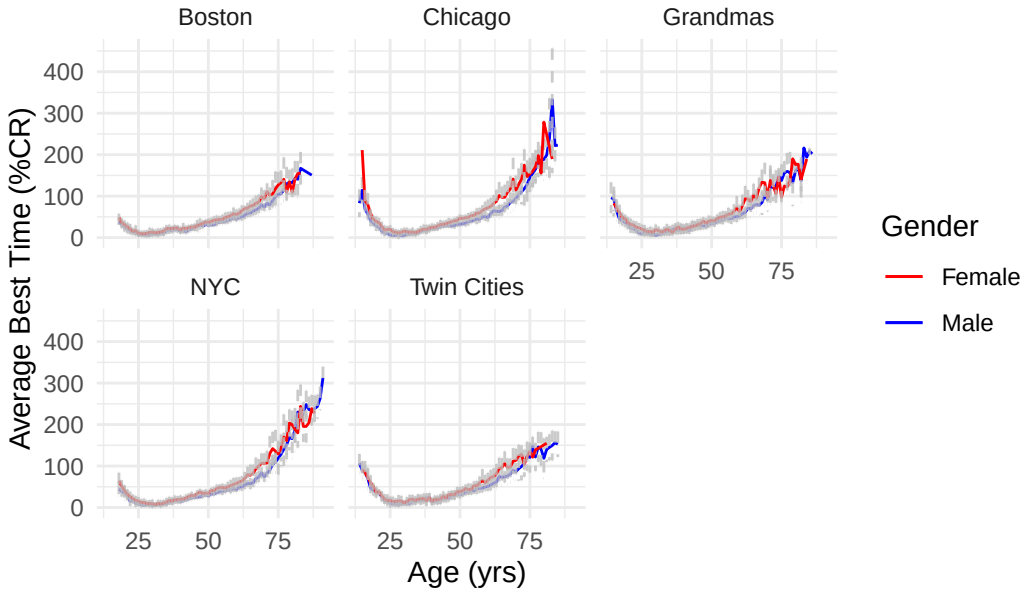


Figure 1 shows a U-shaped pattern in performance across all marathons. Runners reach their fastest times at younger ages and then gradually slow down as they get older. The decline in performance is steeper for females than for males as they age. Additionally, marathons in Chicago and NYC show the most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

Age of Peak Performances

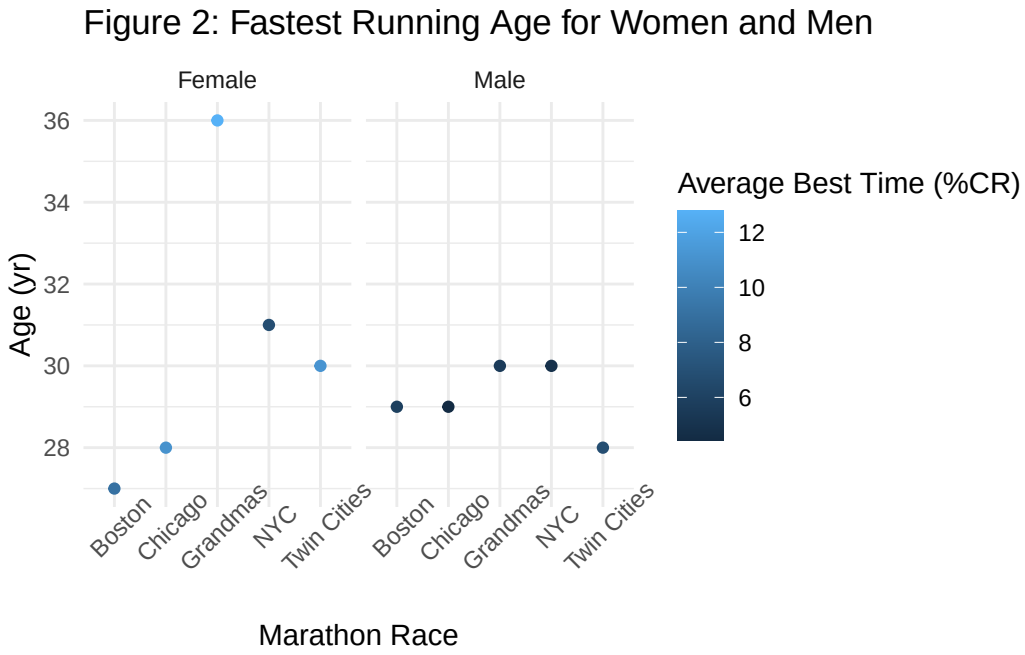


Figure 2 shows that the fastest marathon running age is around 30 years old for both male and females across all five races. However, males show more consistency in their peak running ages across the marathons, with less variation compared to females. For females, the fastest running ages tend to vary 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 slowing becomes significnat, we use a window size of 5 (5 years) to calculate the slope of %CR changing. By looking at the distribution of slopes, we get the median to be around 2 for both male and female, hence we set the threshold to 2. That is, we define the age when slope is greater than 2 to be the age with significantly slowing. At this age, the rate of performance slowing is 2 times the increase in age.

From Table 4, we can observe that the age of significant slowing for male and female is 42 and 41 respectively.

Table 4: Age of significant slowing for women and men n

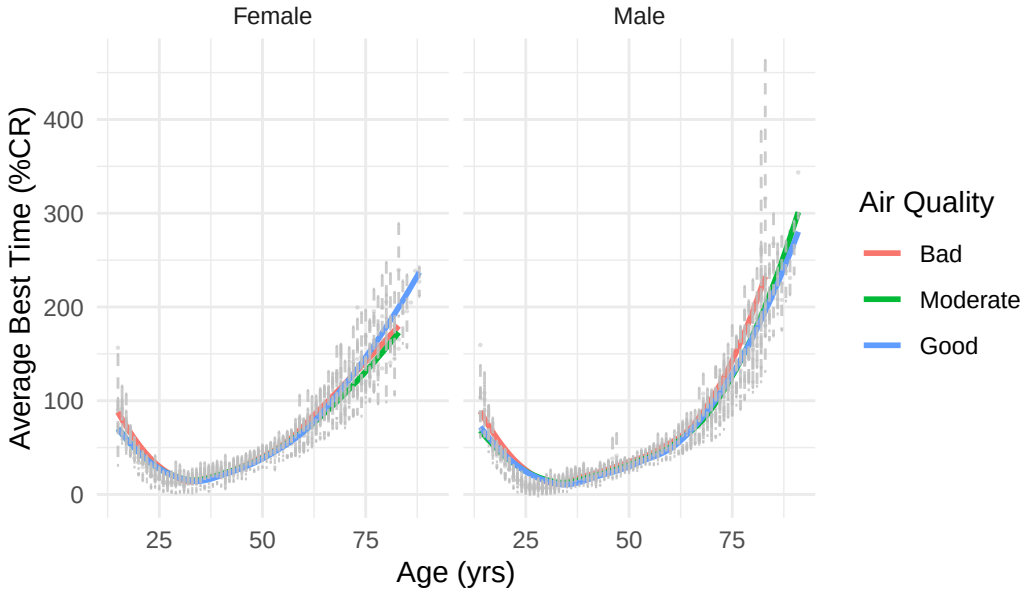
Age of significant slowing	Average %CR	Slope	Sex
42	18.31	2.03	Male
41	22.20	2.11	Female

Impact of environmental conditions on marathon performance

Air Quality Impact

Environmental conditions include weather parameters as well as the air quality parameter (AQI values). We choose AQI estimation taken from duration `8-HR RUN AVG BEGIN HOUR`. We calculated the average AQI values for each marathon in each year and divide them into categories. Based on the distribution of average AQI values, the majority of AQI values lie between 20 to 60. Therefore, categories are defined as Good: AQI < 30, Moderate: AQI 30-50, Bad: AQI > 50. The AQI category is added to the main `project1` dataset.

Figure 3: Effect of Air Quality on Marathon Performances



The U-shaped curve is observed again here in Figure 3. Across both genders, there appears to be minimal separation between the AQI categories when runners are younger (ages 20-40). This suggests that air quality has a limited impact on younger runners, or that the performance

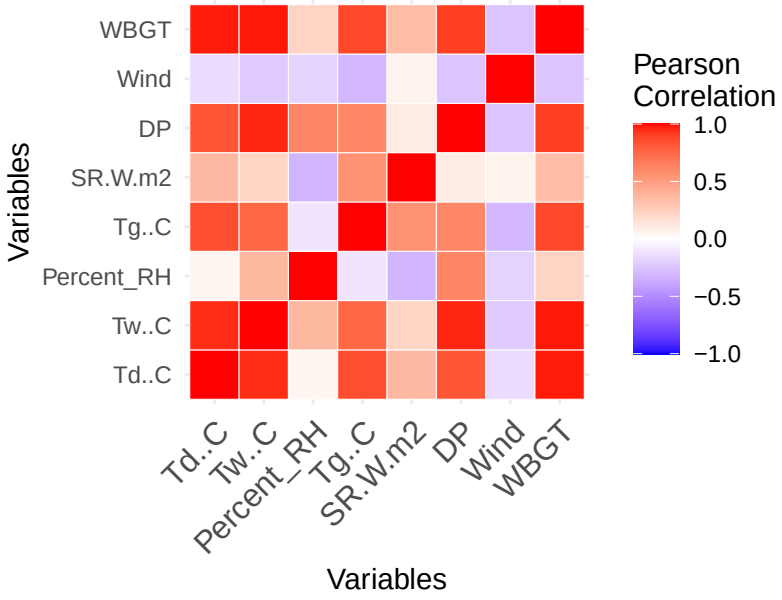
differences are not substantial in this age range. After around age 50, the separation between AQI categories becomes more obvious. For male, we can see that runners performing in "Bad" air quality conditions tend to have higher %CR values compared to those in "Good" air quality conditions. Yet this pattern is not apperant in female runners. Higher variability in performance among older runners is also observed. Overall, air quality might have a slightly greater impact on performance in older marathon runners, and the impact is slightly greater in males than in females.

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 4: Correlation Heatmap



The following plots illustrate the effect of weather parameters (`WBGT`, `Percent_RH`: humidity, `Wind`, and `SR.W.m2`: solar radiation) on marathon performance across age groups and gender. Bins for each weather parameter and age were created by dividing their distribution into quartiles.

Across all four plots, the impact of weather conditions on performance appears minimal, as distributions largely overlap across weather bins within each age group and gender. The primary variations are seen between age groups rather than within the weather parameter bins. While there are some variations in performance spread and median values espeically for older age group, they are not significant enough to indicate strong effects of any particular weather parameter on marathon performance. This suggests that WBGT, humidity, wind, and solar radiation may not have a strong, direct impact on marathon performance.

Figure 5: Effect of WBGT on Marathon Performance

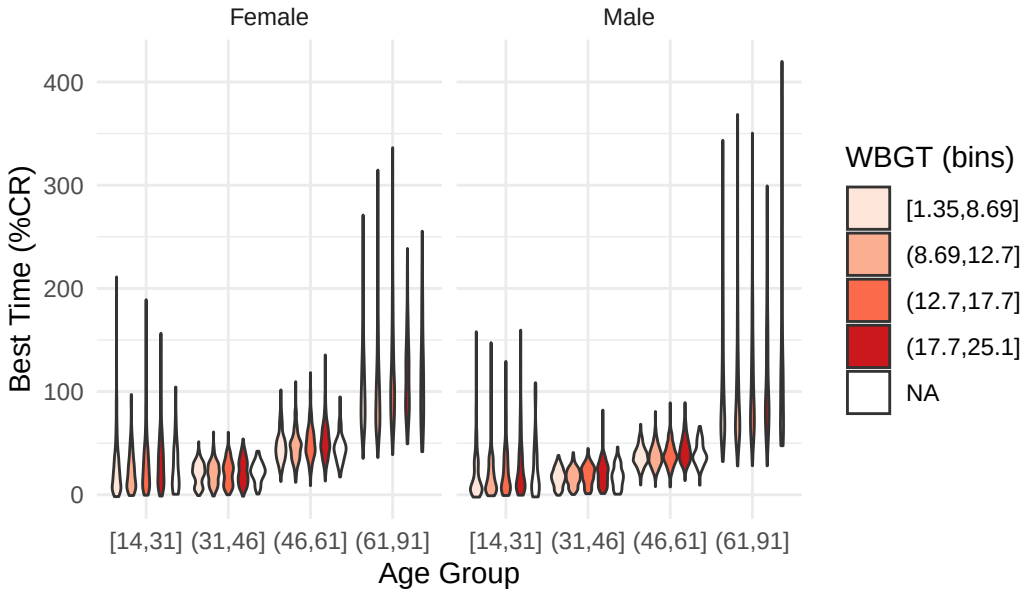


Figure 6: Effect of Humidity on Marathon Performance

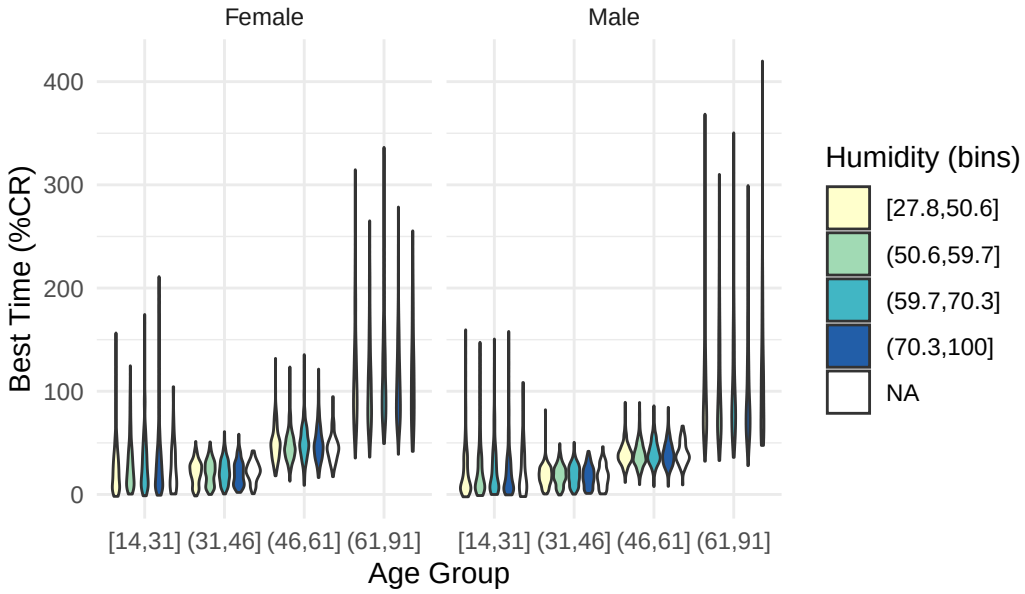


Figure 7: Effect of Solar Radiation on Marathon Performance

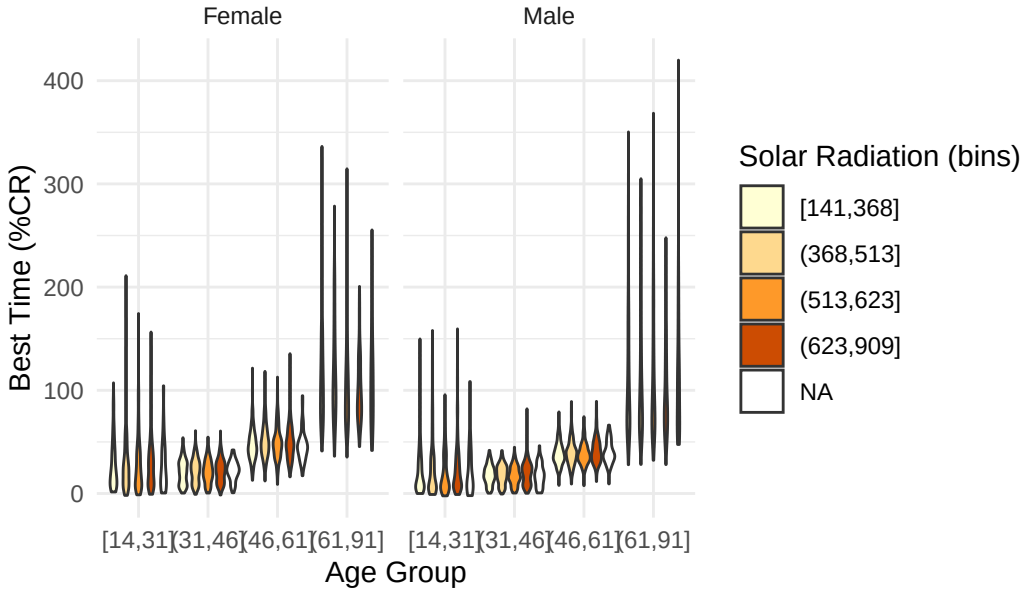
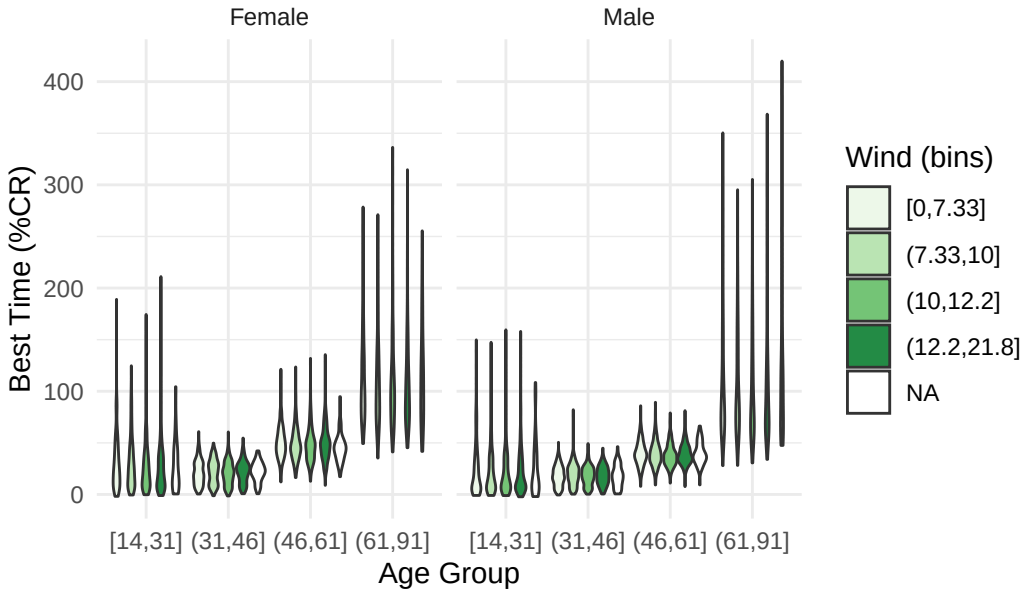


Figure 8: Effect of Wind on Marathon Performance



Regression Analysis

Given the absence of a clear visual impact in previous EDA plots, we fitted a linear model to understand the effect of weather parameters on marathon performance. Table 5 presents only those variables that are statistically significant. Weather parameters like `Percent_RH` (humidity), `SR.W.m2` (solar radiation), `Wind`, and `WBGT` are statistically significant, but their coefficients are small, indicating that their actual impact on performance is limited. The most influential factor in the model is `Age`, which shows a much larger effect on marathon performance. This is consistent with what we observed in the EDA analysis. Also, `Sex` is not significant in this model, suggesting that gender differences do not play a substantial role in this context.

$$
\begin{aligned}
\text{Percent\_CR} = {} & \beta_0 + \beta_1 \cdot \text{Percent\_RH} + \beta_2 \cdot \text{Sex} + \beta_3 \cdot \text{Percent\_RH} \cdot \text{Sex} + \beta_4 \cdot \text{Age} + \beta_5 \cdot \text{Percent\_RH} \cdot \text{Age} \\
& + \beta_6 \cdot \text{SR.W.m2} + \beta_7 \cdot \text{SR.W.m2} \cdot \text{Sex} + \beta_8 \cdot \text{SR.W.m2} \cdot \text{Age} \\
& + \beta_9 \cdot \text{Wind} + \beta_{10} \cdot \text{Wind} \cdot \text{Sex} + \beta_{11} \cdot \text{Wind} \cdot \text{Age} \\
& + \beta_{12} \cdot \text{WBGT} + \beta_{13} \cdot \text{WBGT} \cdot \text{Sex} + \beta_{14} \cdot \text{WBGT} \cdot \text{Age} \\
& + \beta_{15} \cdot \text{aqi\_category} + \beta_{16} \cdot \text{aqi\_category} \cdot \text{Sex} + \beta_{17} \cdot \text{aqi\_category} \cdot \text{Age} \\
& + \epsilon
\end{aligned}
$$

Table 5: Significant variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.1214312	7.6064273	-7.509627	0.0000000
Percent_RH	0.2684858	0.0607778	4.417496	0.0000101
Age	2.3101466	0.1473613	15.676750	0.0000000
SR.W.m2	0.0206351	0.0060227	3.426196	0.0006143
Wind	-0.8427185	0.2286063	-3.686331	0.0002286
WBGT	0.4964138	0.1895523	2.618875	0.0088341
Percent_RH:Age	-0.0080425	0.0011769	-6.833556	0.0000000
Age:SR.W.m2	-0.0006436	0.0001179	-5.459404	0.0000000
Age:Wind	0.0176764	0.0043545	4.059363	0.0000495

Discussion

This report uses multiple exploratory analysis methods such as missing data detection, summary tables and plots along with a linear regression to investigate the relationship between weather conditions and marathon performances, and how it varies across age and gender. The results show that age is the most important factor in marathon performances. Runners' performances peak around 30 and gradually slows down after approximately 40 years old. Weather conditions humidity, solar radiation, Wind and WBGT show statistically significance, however the estimate coefficients are small, leading to very limited impact on the performances. Also, the exploratory data analysis (EDA) and regression results indicate no significant differences in the impact of weather on marathon performance between males and females.

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. The measurement of air quality is another limitation; the '8-HR RUN AVG BEGIN HOUR' variable lacks clarity in terms of its duration. 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

```
# lIbrary used in this analysis
library(dplyr)
library(knitr)
library(kableExtra)
library(visdat)
library(ggplot2)
library(gtsummary)
library(reshape2)
library(car)
library(RColorBrewer)
# 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,
    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,
    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

# Select weather conditions columns
weather_data <- project1 %>%
  dplyr::select(-Sex,-Age,-Percent_CR, -Year)

# Create a heatmap to visualize NA values
vis_miss(weather_data[,-which(names(weather_data) == "marathon_name")],
         sort_miss = TRUE) +
  theme(axis.text.x = element_text(angle = 90, size = 10))
# 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 = "3cm") %>%
  column_spec(3, width = "3cm") %>%
  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(title="Figure 1: Effect of Age on Marathon Performances",
       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) %>%
```

```r
  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)
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(title="Figure 2: Fastest Running Age for Women and Men",
       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)

# Find the median slopes for male and female to determine a threshold
median_slope_female <- median(age_data_female_slopes$slope, na.rm = TRUE)
median_slope_male <- median(age_data_male_slopes$slope, na.rm = TRUE)

# Set a threshold to detect slowing
slope_threshold <- 2

# Identify the first age where the slope exceeds the threshold
slowing_point_male <- age_data_male_slopes %>%
  filter(slope > slope_threshold) %>%
  slice(1)
slowing_point_female <- age_data_female_slopes %>%
  filter(slope > 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 n",
        longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down"))
```

```r
slowing_point_combined
# Calculate average aqi values for each year in each marathon
aqi_average <- aqi_values %>%
  filter(sample_duration == "8-HR RUN AVG BEGIN HOUR") %>%
  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")))

# Custom categorization
aqi_average <- aqi_average %>%
  mutate(aqi_category = case_when(
    mean_aqi < 30 ~ "Good",
    mean_aqi <= 50 ~ "Moderate",
    TRUE ~ "Bad"
  ))

# Join `project1` with `aqi_acerage` based on `marathon` and `Year`
project1 <- project1 %>%
  left_join(aqi_average %>% dplyr::select(-c(date_local,mean_aqi)),
            by = c("marathon_name" = "marathon", "Year" = "Year"))
# Create a data frame summarising %CR by air quality, sex and age
air_quality_data <- project1 %>%
  group_by(aqi_category, Sex, Age) %>%
  summarise(performance = mean(Percent_CR, na.rm = TRUE),
            sd_CR = sd(Percent_CR, na.rm=TRUE))

# Reorder all levels of aqi_category
air_quality_data$aqi_category <- factor(air_quality_data$aqi_category,
                                        levels = c("Bad", "Moderate", "Good"))

# VIsualize air quality across age and gender
air_quality_plot <- ggplot(data=air_quality_data)+
  geom_point(aes(x=Age, y=performance), color='grey',alpha=0.5, size=0.1)+
  geom_smooth(data=air_quality_data, aes(x = Age, y = performance, color=aqi_category),
              method = "loess", se = F) +
  geom_errorbar(aes(x=Age, ymin = performance -sd_CR,
                    ymax = performance +sd_CR),
                color = "grey", width = 0.5, linetype="dashed", alpha=0.8) +
```

```
  facet_wrap(~Sex)+
  theme_minimal()+
  labs(title="Figure 3: Effect of Air Quality on Marathon Performances",
       color="Air Quality",
       x="Age (yrs)", 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(title = "Figure 4: Correlation Heatmap",
       x = "Variables", y = "Variables")

cor_heatmap
# Create quantile-based bins for age, WBGT,SR.W.m2, Wind, humidity
project1$Age_Group <- cut(project1$Age, breaks = quantile(project1$Age,
                                                          probs = seq(0, 1, 0.25),
                                                          na.rm=TRUE),
                          include.lowest = TRUE)
project1$WBGT_bin <- cut(project1$WBGT, breaks = quantile(project1$WBGT,
                                                          probs = seq(0, 1, 0.25),
                                                          na.rm=TRUE),
                         include.lowest = TRUE)
project1$SR_bin <- cut(project1$SR.W.m2, breaks = quantile(project1$SR.W.m2,
                                                           probs = seq(0, 1, 0.25),
```

```
                                                        na.rm=TRUE),
                        include.lowest = TRUE)
project1$Wind_bin <- cut(project1$Wind, breaks = quantile(project1$Wind,
                                                          probs = seq(0, 1, 0.25),
                                                          na.rm=TRUE),
                         include.lowest = TRUE)
project1$RH_bin <- cut(project1$Percent_RH, breaks = quantile(project1$Percent_RH,
                                                              probs = seq(0, 1, 0.25),
                                                              na.rm=TRUE),
                       include.lowest = TRUE)

# Plot for WBGT impact on performances
WBGT_plot <- ggplot(project1, aes(x = Age_Group, y = Percent_CR, fill = WBGT_bin)) +
  geom_violin() +
  facet_wrap(~ Sex) +
  labs(
    title = "Figure 5: Effect of WBGT on Marathon Performance",
    x = "Age Group",
    y = "Best Time (%CR)",
    fill = "WBGT (bins)"
  ) +
  scale_fill_brewer(palette = "Reds") +
  theme_minimal()
WBGT_plot


# Plot for humidity impact on performances
humidity_plot <- ggplot(project1, aes(x = Age_Group, y = Percent_CR, fill = RH_bin)) +
  geom_violin() +
  facet_wrap(~ Sex) +
  labs(
    title = "Figure 6: Effect of Humidity on Marathon Performance",
    x = "Age Group",
    y = "Best Time (%CR)",
    fill = "Humidity (bins)"
  ) +
  scale_fill_brewer(palette = "YlGnBu") +
  theme_minimal()
humidity_plot
# Plot for solar radiation impact on performances
solar_plot <- ggplot(project1, aes(x = Age_Group, y = Percent_CR, fill = SR_bin)) +
```

```
  geom_violin() +
  facet_wrap(~ Sex) +
  labs(
    title = "Figure 7: Effect of Solar Radiation on Marathon Performance",
    x = "Age Group",
    y = "Best Time (%CR)",
    fill = "Solar Radiation (bins)"
  ) +
  scale_fill_brewer(palette = "YlOrBr") +
  theme_minimal()
solar_plot
# Plot for Wind impact on performances
wind_plot <- ggplot(project1, aes(x = Age_Group, y = Percent_CR, fill = Wind_bin)) +
  geom_violin() +
  facet_wrap(~ Sex) +
  labs(
    title = "Figure 8: Effect of Wind on Marathon Performance",
    x = "Age Group",
    y = "Best Time (%CR)",
    fill = "Wind (bins)"
  ) +
  scale_fill_brewer(palette = "Greens") +
  theme_minimal()
wind_plot
# Fit a linear model
model <- lm(Percent_CR ~ Percent_RH * Sex + Percent_RH * Age +
                         SR.W.m2 * Sex + SR.W.m2 * Age +
                         Wind * Sex + Wind * Age +
                         WBGT * Sex + WBGT * Age +
                         aqi_category * Sex + aqi_category * Age, data = project1)

# Collect model summary
model_summary <- summary(model)

# Extract coefficients as a data frame
coef_df <- as.data.frame(model_summary$coefficients)

# Change column names
colnames(coef_df) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

# Filter only significant variables (e.g., p < 0.05)
```

```r
significant_vars <- coef_df[coef_df$`Pr(>|t|)` < 0.05, ]
significant_vars%>%
  kable(booktabs = TRUE, caption = "Significant variables", longtable = TRUE,
        linesep = "") %>%
  kable_styling(font_size = 8,
latex_options = c("repeat_header", "HOLD_position", "scale_down"))
```