Exploring the Impact of Environment and Weather Conditions on Marathon Performance

Tianna Chan

2024-10-06

Abstract

This project explores the impact of environmental factors—including temperature, humidity, solar radiation, wind, and air quality—on marathon performance across gender and age. Using data from major U.S. marathons (1993-2016), combined with weather and air quality information, the relationships between these factors and race times were examined. Results reveals significant differences based on gender and age, which also applies to air quality impacts in the measures of PM2.5 and ozone levels. Additionally, relative humidity and wet bulb globe temperature are the most influential weather factors affecting performance, with effects vary by age. These findings offer valuable insights for athletes and coaches to optimize race-day strategies based on environmental conditions and individual characteristics.

Introduction

Previous studies have shown aerobic performance declines even under just modest hyperthermia in hot environments [1]. Specifically, in marathon running, higher temperatures are associated with slower race times [2]. Additionally, older adults may struggle more in hot and humid conditions due to a reduced ability to dissipate heat effectively [3]. Sex differences in endurance running have also been observed, with male and female exhibiting distinct physiological and thermoregulatory responses under environmental conditions [4].

This project aims to explore different environmental factors that impact marathon performance between gender and across age. This includes not only temperature, but humidity, solar radiation, wind, and air quality on a race date. Understanding the associations provides a more holistic comprehension on how environmental conditions influence performance in endurance events such as marathon. It could also be beneficial for athletes or coaches in forming their targeted strategies based on their own characteristics or weather.

We will start with data preprocessing and begin the exploratory, by first examining the age and sex effects on marathon performance. Then, we will factor in the air quality information for PM2.5 and Ozone level. Lastly, weather parameters will be investigated and a discussion on the results is included. All explorations will be presented through plots, tables, and regression models.

The Data and Data Preprocessing

Project 1 & Course Record Data

The `project1` data set contains info and results for 11564 participants (14-85 years) in the marathon races at Boston, Chicago, New York City, Twin Cities, and Grandmas from 1993 to 2016, with their percent off course record. It also includes 9 weather parameters, including the dry bulb and wet bulb temperature ($^{\circ}C$) , percent relative humidity (%) , black globe temperature ($^{\circ}C$) , solar radiation (W/m^2), dew point ($^{\circ}C$), wind speed (km/hr), wet bulb globe temperature ($^{\circ}C$), and a flag based on the wet bulb globe temperature and risk of heat illness. The Wet Bulb Globe Temperature is the weighted average of the dry bulb, wet bulb, and globe temperature, which the wet bulb factors in humidity and globe temperature factors factors in solar radiation.

Column names were cleaned for readability and simplicity. Re-coding was done to some columns to help ease the process of comprehending and merging later- We re-coded and factorized the Race column into Boston, Chicago, New York City, Twin Cities, and Grandmas; And the Sex column into Male and Female. Additionally, Flag were factorized with White (the lowest risk group) as the reference level for future analyses.

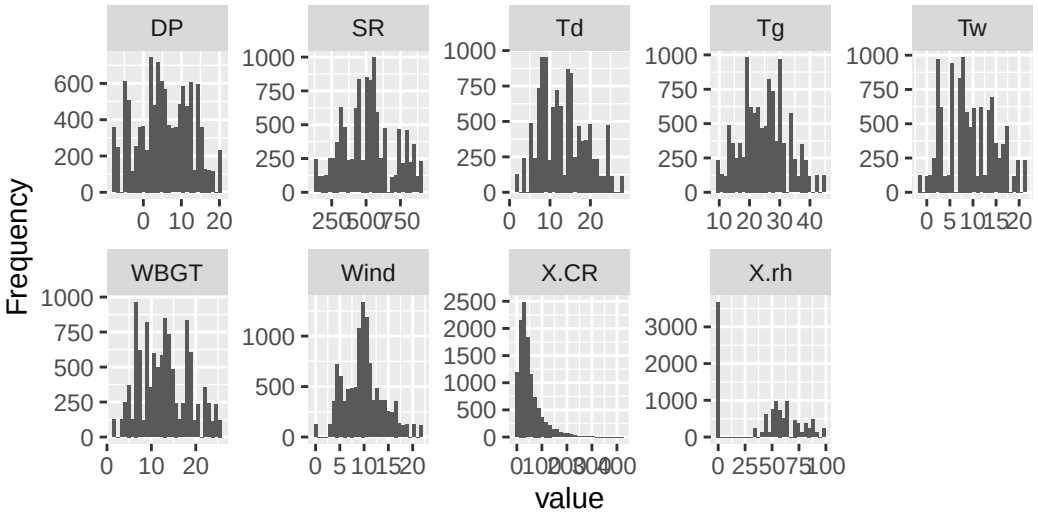

Figure 1: Histogram for Numeric Project 1 Data

As histograms (**Figure 1**) were plotted for all the numeric values in the data, we observe that the percent off course record was heavily right-skewed, and the percent relative humidity has a big spike around zero, indicating potential data quality. Upon further investigation, there were a portion of data that recorded relative humidity in the scale of 0 to 1, while others were 0% to 100%. All those scaled 0 to 1 were later multiplied by 100 to match the rest, the distribution after the correction looks normal.

We observed the missing data plot by `vis_miss()`. There was a consistent 4% of missing data for the flag and weather measurement columns, which indicates that weather was not measured at these races. After some exploration, we found that all weather data were missing at races in Chicago, New York City, and Twin Cities at 2011, and Grandma's at 2012. These weather variables were missing because they are not measured in that particular year/race, but we do not have knowledge on the exact reason. Therefore, the probability of the weather parameters being missing do not depends on any environmental or course information, which should then be a case of Missing Completely at Random (MCAR).

The project also utilized the `course_record` data that included the course record for each gender at each race and year. No missingness was found in the data.

Air Quality Data

To further assess the environmental impact, AQI data was obtained using the R package RAQSAPI, grabbing data from the US Environmental Protection Agency's API. The code was provided and available at `aqi.R`. The resulted data set includes the arithmetic mean values for ozone level in parts per million, and the PM2.5 in Micrograms/cubic meter (LC). The sample duration for the data includes 1 hour, 24 hour, 24-hour block average, and 8-hour run average begin hour. The marathon names and dates were recoded to facilitate the merging of the data later, and the function `year()` was used to obtain the year only.

Based on the missingness graph, the AQI column has 16% missing values. We found that all data with sample duration = 1 hour has the AQI column missing. This is then Missing At Random (MAR) since it only depends on this observed column.

According to the US Environmental Protection Agency, 88101 and 88502 are both used to report daily Air Quality Index values. The difference is that 88101 include both manually operated Federal Reference Methods (FRMs) and automated Federal Equivalent Methods (FEMs), but 88502 are "FRM-like" @EPA. Based on this information, only the 88101 data was taken for simplicity. We created a data summary by taking the average of the arithmetic mean. However, note that there were incomplete data - it contains missing PM2.5 values from most of the early years and a few in later years. It is possible that the PM2.5 values were not collected until later years. This could be a Missing At Random (MAR) case because it only depends on observed variable `year`.

Merging the Data

The course record data was merged to project 1 data using a `left_join()`. Afterwards, the record was then used to calculate a new variable `Time` for each participant, with the equation Time (hours) $= \frac{\text{CR} \times \left(1 + \frac{\%CR}{100}\right)}{3600}$ that adjusts for the percent off course record and convert the final unit to hours. Next, the AQI data was also joined using `left_join()`. **Figure 2** shows the correlations between each variable after merging the data. High correlations were seen between Wet bulb globe termperature (WGBT) and the three other temperaturest (Td, Tw, Tg), and also dew point (DP). Thus, we will only use WGBT and omit the others in future analyses. We also see high correlations between the percent of course record and age (0.7), which we will dive deeper and quantify these relations between the variables in the later sections.

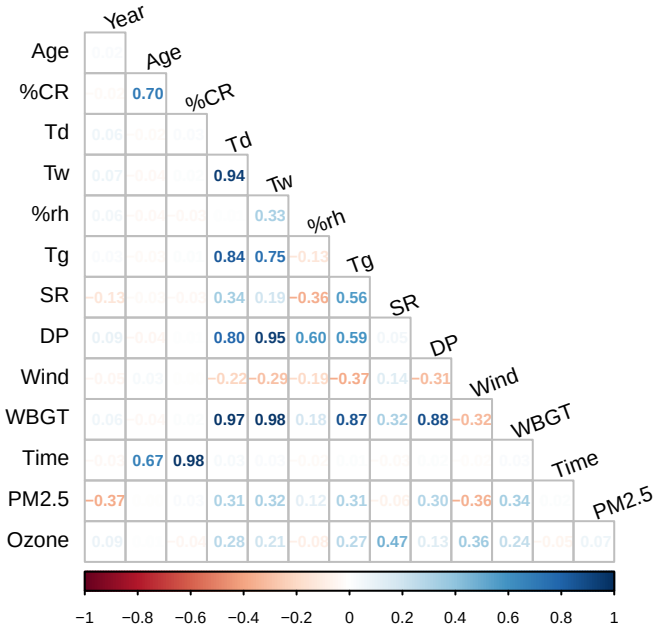

Figure 2: Correlation plot

Effects of increasing age on marathon performance in men and women

Figure 3 below shows a violin plot of time and age for men and women, in 5-year age intervals. The violin plots illustrates the distribution of course time within each age groups. The distribution of course times for women is slightly higher than for men across all age groups, suggesting that, on average, women run slightly slower than men. In overall, there seem to be a near-identical trend for men and women - they both improve their times as they approach the age group (25,30], and after this point, course times gradually increase, indicating slower performance with advancing age. It is hard to tell from the graph which sex's course time increase is at a steeper rate. Thus, we will later fit a model to quantify the changes.

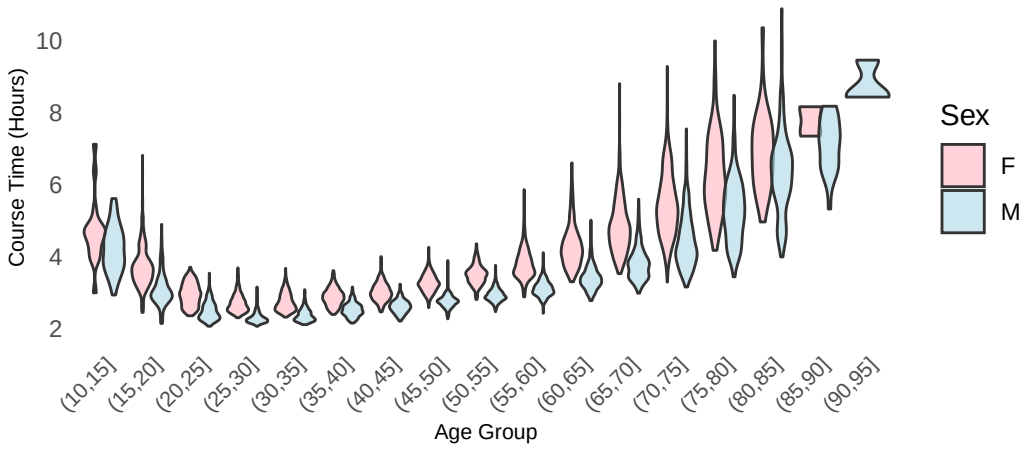

Figure 3: Course Time vs Age by Sex

This table shows the summary statistics for age and course time for men and women. It is obvious that men (N = 6,112) had a larger sample size than women (N = 5,452). In average, the men population is also older than the female population. We again see that men have a shorter course time on average.

Table 1: Age and Time Summary by Sex

Characteristic	**F** N = 5,452	**M** N = 6,112
Age	45 (30, 59)	48 (32, 64)
Time	3.55 (2.87, 3.91)	3.17 (2.49, 3.46)
[1] Mean (Q1, Q3)		

Regression models were fit to quantify the course time by age for each sex. We included age-squared in this model to allow for a non-linear relationship between age and course time, as seen in previous graphs. The final models are as follows:

$$E[\text{Time (Female)}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Age}^2 (i)$$
$$E[\text{Time (Male)}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Age}^2 (ii)$$

Table 2: Linear Regression Model Summary

	Female				Male			
	Estimate	Std Error	t value	p value	Estimate	Std Error	t value	p value
Intercept	5.265	0.045	117.675	0	4.746	0.037	128.401	0

Age	-0.138	0.002	-65.544	0	-0.128	0.002	-76.744	0
Age Squared	0.002	0.000	86.009	0	0.002	0.000	101.313	0

Based on the results below, both Age and Age^2 are highly significantly associated on time ($p < 0.001$), with Age negatively associated and Age^2 positively associated.

The intercept for females is 5.265, and 4.746 for males. This suggests that, hypothetically at age zero, the estimated course time would be slightly higher for females than for males.

The coefficient for Age is -0.138 for females and -0.128 for males, indicating a strong negative relationship between age and course time in the younger age range. In other words, as age increases, course time decreases up to a certain point. The slightly more negative value for females suggests that females initially show a slightly faster improvement in course time with each year of age compared to males.

The coefficient for Age^2 being 0.002 for both females and males indicates a non-linear relationship between age and course time, which the course time begins to increase as it passes certain age, reflecting the U-shaped trend as plotted before. The merely same coefficient means that the rate at which performance slows down after certain age is very similar for both sexes.

In conclusion, females may have a slightly higher baseline course time than males, and they generally experience a faster improvement in their younger ages compared to males, but then have similar rates of decline in performance after reaching their optimal age.

Impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

The plots below show the relationship between environmental conditions (Ozone and PM2.5 levels) and marathon performance (course time) across different age groups of 10 by gender. Each panel represents a specific age group, with Ozone or PM2.5 levels on the x-axis and course time (in hours) on the y-axis. The points corresponds to individual data points, with colors indicating the sex. Smoothing lines for each gender were plotted using `geom_smooth`, which highlights the trends within each age group, indicating how variations in environmental conditions might impact performance differently for males and females. This visualization allows us to examine whether higher levels of Ozone or PM2.5 are associated with slower course times, and if these effects differ across ages and between genders.

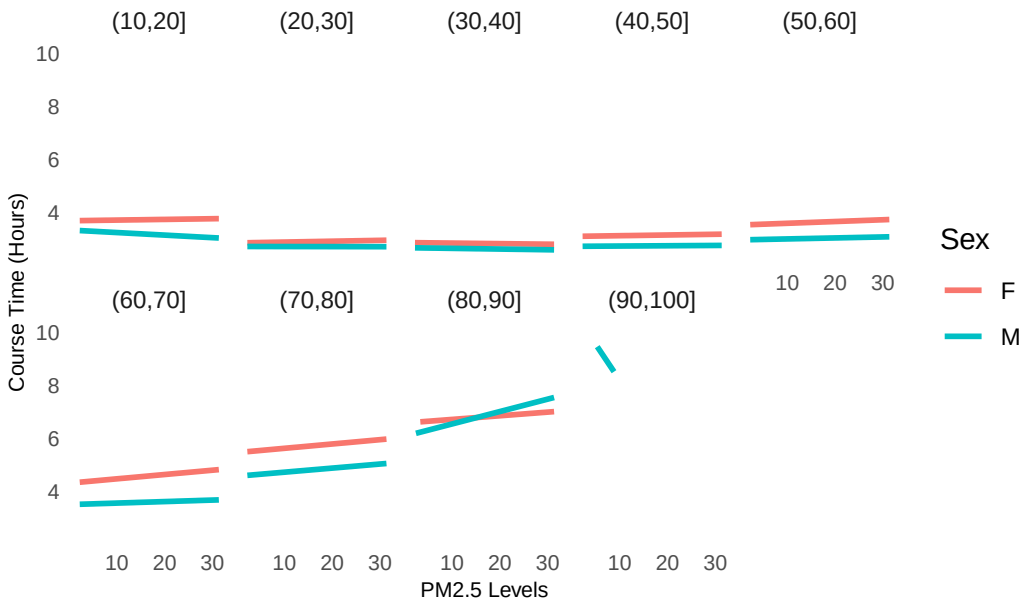

Figure 4: Impact of PM2.5 on Marathon Performance by Age Group and Sex

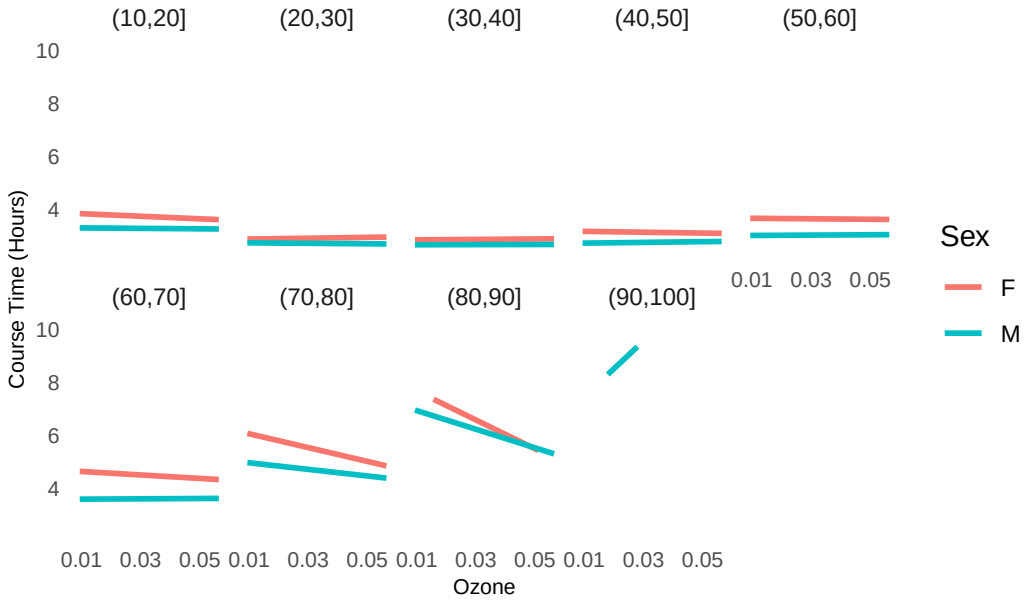

Figure 5: Impact of Ozone on Marathon Performance by Age Group and Sex

To quantify the impact of environmental conditions on marathon performance, we fit a regression model building on the previous model, with age, sex, PM2.5 and Ozone. Based on previous plots, we see a more significant impact on environmental conditions at 60 years and

more. Thus, we further added an age group indicator variable of whether age is 60 or above.

The full model: $E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM} + \beta_3 \cdot \text{Age}^2 + \beta_4 \cdot \text{PM2.5} + \beta_5 \cdot \text{60andAbove} + \beta_6 \cdot \text{Ozone} + \beta_7 \cdot \text{Age} \cdot \text{SexM} + \beta_8 \cdot \text{Age} \cdot \text{PM2.5} + \beta_9 \cdot \text{60andAbove} \cdot \text{PM2.5} + \beta_{10} \cdot \text{SexM} \cdot \text{PM2.5} + \beta_{11} \cdot \text{Age} \cdot \text{Ozone} + \beta_{12} \cdot \text{60andAbove} \cdot \text{Ozone} + \beta_{13} \cdot \text{SexM} \cdot \text{Ozone}$

Interaction is included between sex and age as we saw differences between the two sexes from previous results. We also included the two-way interactions of Age/Age group/Sex with PM2.5/Ozone because of the observable trends from plots above. Then, backward selection was used to find the parsimonial model using AIC. The Age $\times$ Ozone interaction term was dropped in the backward selection process, which indicates that ozone does not appear to have a varying effect on course time across different ages, but the age grouping of being aged 60 and above is still relevant.

The final backward selection model: $E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM} + \beta_3 \cdot \text{Age}^2 + \beta_4 \cdot \text{PM2.5} + \beta_5 \cdot \text{60andAbove} + \beta_6 \cdot \text{Ozone} + \beta_7 \cdot \text{Age} \cdot \text{SexM} + \beta_8 \cdot \text{Age} \cdot \text{PM2.5} + \beta_9 \cdot \text{60andAbove} \cdot \text{PM2.5} + \beta_{10} \cdot \text{SexM} \cdot \text{PM2.5} + \beta_{11} \cdot \text{60andAbove} \cdot \text{Ozone} + \beta_{12} \cdot \text{SexM} \cdot \text{Ozone}$

Table 3: Linear Regression Model Summary

Variable	**Estimate**	**Std. Error**	**t value**	**p value**
(Intercept)	5.268	0.046	114.387	0.000
Age	-0.139	0.002	-80.448	0.000
SexM	-0.120	0.037	-3.222	0.001
I(Age^2)	0.002	0.000	101.481	0.000
PM2.5	-0.009	0.003	-3.415	0.001
AgeGroup60andAbove	-0.089	0.042	-2.130	0.033
Ozone	-1.980	0.727	-2.724	0.006
Age:SexM	-0.010	0.001	-19.937	0.000
Age:PM2.5	0.000	0.000	4.716	0.000
PM2.5:AgeGroup60andAbove	0.006	0.003	2.209	0.027
SexM:PM2.5	-0.003	0.001	-1.868	0.062
AgeGroup60andAbove:Ozone	-10.426	1.063	-9.809	0.000
SexM:Ozone	1.831	0.946	1.936	0.053

The final model by backward selection kept both PM2.5 and Ozone and most of their two-way interactions with age, age group and sex. All of them are statistically significant except Sex & PM2.5 and Sex & Ozone (with marginal significance). Thus, this means that both environmental conditions were able to explain course time.

For PM2.5, $\beta_4 = -0.009$ means that holding all else constant, each unit increase in PM2.5 is associated with a 0.009-hour (32.4 seconds) decrease in course time. This is counter-intuitive, but its effect is modified by the interaction with age, age group, and sex. Coefficient of $\beta_8 = 0.0002933$ means that increase in age would make the effect less negative, and when they

are 60 and above, there is an additional $\beta_9 = 0.006$ hours (21.6 seconds) increase in course time. This would make the whole effect positive. Coefficient of $\beta_10 = -0.003$ indicates that for males, each unit increase in PM2.5 reduces course time by an additional 0.003 hours (10.8 seconds) compared to female. In other words, the negative impact of PM2.5 on performance becomes more pronounced as age increases and when the age is above 60, but is less negative for males in overall.

For ozone levels, $\beta_6 = -1.980$ means that for every one unit increase in Ozone, the course time decreases by 1.980 hours. Thus, by conversion, for every 0.01 unit increase in Ozone, the course time decreases by 0.0198 hours (1.188 minutes). This effect is also modified by the interaction with age group, coefficient of $\beta_{12} = -10.425$, and the interaction with sex, coefficient of $\beta_{13} = 1.831$. This means that when individuals are aged 60 and above, the effect of Ozone is even more pronounced, with an additional decrease of 10.426 hours in course time per unit increase in Ozone (0.10426 hours/ 6.2556 minutes for every 0.01 unit increase in Ozone). And for males, each unit increase in Ozone is associated with an additional increase in course time of 1.831 hours compared to females (0.01831 hours/ 1.0986 minutes for every 0.01 unit increase in Ozone). This suggests that comparing with females, males may be more negatively affected as ozone levels are higher, and the effect is stronger when aged 60 and above.

Weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

The graph below illustrates the relationship of WBGT, flag, and the marathon performance on course time. We can almost draw vertical lines to pinpoint the appearance of new flag colors at certain time points. To elaborate, only whites and one green had an average course time <3.12, then only whites and greens had an average course time <3.28, and only whites, greens, and yellows < 3.38. This suggests that larger WBGT may be associated with a longer course time.

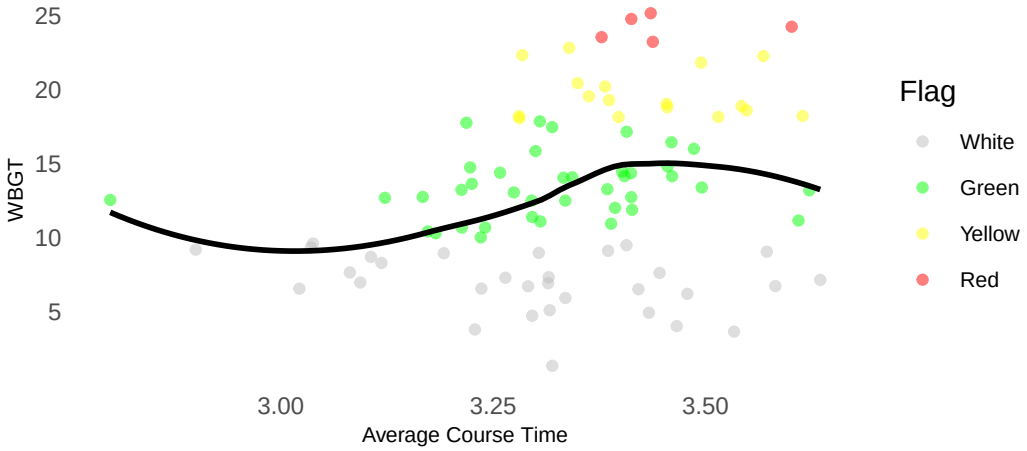

Figure 6: Impact of WBGT on Marathon Performance

The plots below display the relationship between each weather parameter and course time by different age quartiles (Q1, Q2, Q3, and Q4). Each colored line refer to the different age quartiles, and the grey line refers to the overall.

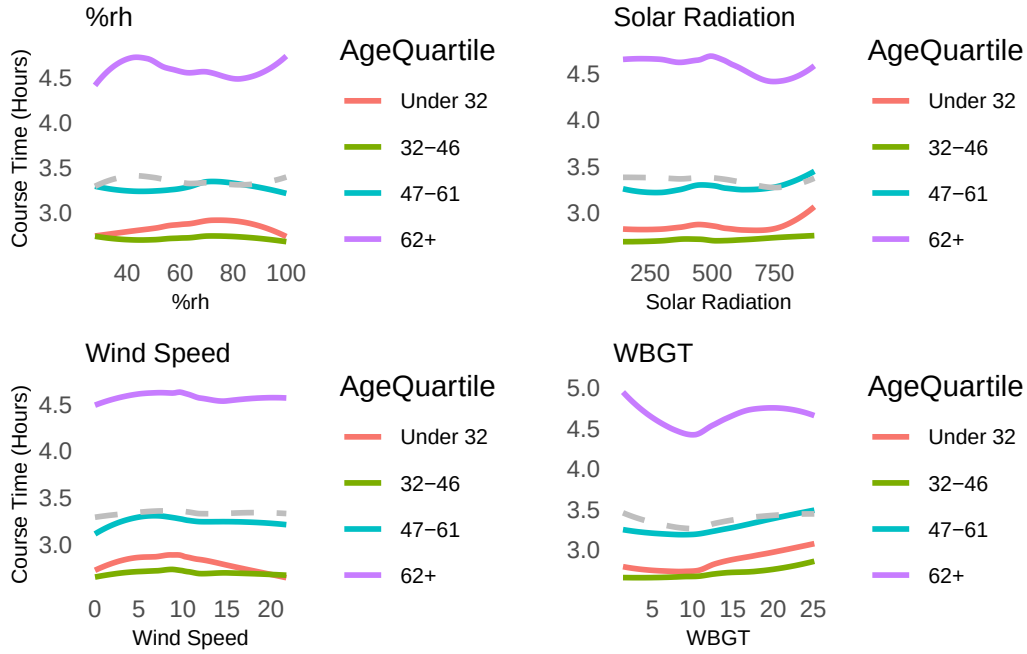

Figure 7: Impact of Weather on Marathon Performance

From the plots above, we found that the weather parameters tend to have a greater impact on the oldest age group (62+), since there are larger variations shown in humidity, solar radiation,

and WBGT. Specifically, there is a slight increase of course time with higher humidity levels but it is non-linear. For solar radiation, it is stable until a certain point, it decreases and increases, though it does not return to the original level. For WBGT, it decreases initially and increases after a certain point, which may indicate potential sensitivity to temperature and humidity combined. For the other age groups, an increasing trend was observed. The effect of wind speed is relatively stable for the oldest age group as well as other age groups.

The other age groups generally exhibit similar patterns. They have a relatively stable course times across different humidity levels and wind speed, indicating that these might be less impactful on course time for younger participants. However, for solar radiation and WBGT, we can see an obvious increasing trend as these parameters increase.

Upon creating the same plot by sex, we see almost the same trend among female and male.

To quantify, we will fit a model with course time regressed on the weather variables and age. Sex was not included for simplicity, since seemingly indifference result was found between both sexes. Then, backward selection was used to find the smallest best model. **Table 8** below shows the coefficients of the final model by backward selection. Because non-linearity is observed from previous plots, we included squared terms for relative humidity, solar radiation, and WBGT.

The initial model: $E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \%\text{rh} + \beta_3 \cdot \text{SR} + \beta_4 \cdot \text{Wind} + \beta_5 \cdot \text{WBGT} + \beta_6 \cdot \%\text{rh}^2 + \beta_7 \cdot \text{SR}^2 + \beta_8 \cdot \text{WBGT}^2 + \beta_9 \cdot \text{Age} \cdot \%\text{rh} + \beta_{10} \cdot \text{Age} \cdot \text{SR} + \beta_{11} \cdot \text{Age} \cdot \text{Wind} + \beta_{12} \cdot \text{Age} \cdot \text{WBGT} + \beta_{13} \cdot \text{Age} \cdot \%\text{rh}^2 + \beta_{14} \cdot \text{Age} \cdot \text{SR}^2 + \beta_{15} \cdot \text{Age} \cdot \text{Wind}^2 + \beta_{16} \cdot \text{Age} \cdot \text{WBGT}^2$

The final model: $E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \%\text{rh} + \beta_3 \cdot \text{Wind} + \beta_4 \cdot \text{WBGT} + \beta_5 \cdot \%\text{rh}^2 + \beta_6 \cdot \text{SR}^2 + \beta_7 \cdot \text{WBGT}^2 + \beta_8 \cdot \text{Age} \cdot \%\text{rh} + \beta_9 \cdot \text{Age} \cdot \text{Wind} + \beta_{10} \cdot \text{Age} \cdot \text{WBGT} + \beta_{11} \cdot \text{Age} \cdot \%\text{rh}^2 + \beta_{12} \cdot \text{Age} \cdot \text{SR}^2$

Table 4: Linear Regression Model Summary: Backward Selection

Variable	**Estimate**	**Std. Error**	**t value**	**Significance**
(Intercept)	0.1239	0.2675	0.4633	0.6432
Age	0.0718	0.0053	13.5694	0.0000
`%rh`	0.0367	0.0079	4.6392	0.0000
Wind	-0.0114	0.0053	-2.1417	0.0322
WBGT	-0.0131	0.0074	-1.7711	0.0766
I(`%rh`^2)	-0.0002	0.0001	-3.7155	0.0002
I(SR^2)	0.0000	0.0000	7.0580	0.0000
I(WBGT^2)	0.0008	0.0002	3.8121	0.0001
Age:`%rh`	-0.0009	0.0002	-5.4191	0.0000
Age:Wind	0.0002	0.0001	2.1387	0.0325
Age:WBGT	0.0002	0.0001	1.9727	0.0486
Age:I(`%rh`^2)	0.0000	0.0000	4.2140	0.0000
Age:I(SR^2)	0.0000	0.0000	-9.4554	0.0000

Looking at the estimates above, it is clear that almost all of the weather parameters considered impacts marathon performance in some level.

The most impactful weather parameters include the percent relative humidity (both linear, squared) and WBGT (both linear and squared), with their effects vary by age, since their interaction with age is significant. Solar radiation also has a highly significant non-linear effect on course time, as well as its interaction with age. Wind also provides some benefit with significant results.

Discussion

This project explored the impact of environmental and weather conditions such as PM2.5, ozone levels, temperature, and humidity on marathon performance across different age group and genders. The findings revealed significant associations between some of these variables and race times, which varied by age and sex.

The data used in this analysis has a wide range of participants and races under different conditions. However, there are limitations that must be acknowledged. Specifically, the PM2.5 and Ozone levels were calculated by taking the average of all sample duration (e.g., 1-hour, 8-hour, and 24-hour measurements), which may not fully capture the variability of air quality experienced by participants throughout the marathon. The age groupings used in the analyses were also considered in broad categories, but there may be finer age-related effects with categories. Future work could look for more specific age groups.

In conclusion, this analysis demonstrates that PM2.5, ozone levels, relative humidity, WBGT, and solar radiation were significantly associated with marathon performance. Wind also shows some beneficial effects and significantly improving performance. The environmental conditions vary by age and less so by sex, while the weather conditions only vary by age.

References

1. 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.
2. 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.
3. Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.
4. 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.

Code Appendix

```
# Load libraries
library(readr)
library(tidyverse)
set.seed(123456)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(visdat)
library(gtsummary)
library(patchwork)
library(DataExplorer)
library(gridExtra)
# Import data sets
aqi_values <- read_csv("aqi_values.csv")
course_record <- read_csv("course_record.csv")
project1 <- read_csv("project1.csv")

###### DATA PREPROCESSING ######

## Project 1 Data

# Project 1 - Data Quality
# str(project1)

# Project 1 - Rename column names
colnames(project1) <- c("Race", "Year", "Sex", "Flag", "Age", "%CR", "Td",
 ↪  "Tw", "%rh", "Tg", "SR", "DP", "Wind", "WBGT")

# Project 1 - Recode columns
project1 <- project1 %>%
  mutate(Race = case_when(
    Race == 0 ~ "B",
    Race == 1 ~ "C",
    Race == 2 ~ "NY",
    Race == 3 ~ "TC",
    Race == 4 ~ "D",
    ),
    Sex = case_when(
```

```
    Sex == 0 ~ "F",
    Sex == 1 ~ "M"
    )
    ) %>%
  mutate(across(c(Race, Sex, Flag),
         as.factor))

# Project 1 - Set white flag as reference
project1$Flag <- factor(project1$Flag, levels = c("White", "Green", "Yellow",
 ↪  "Red"))

# FIGURE 1: Project 1 - Plot histogram
hist_data <- project1 %>% select(-c(Age, Year))
plot_histogram(hist_data, ncol = 5L)

# Fix relative humidity data quality issue
project1$`%rh` <- case_when(project1$`%rh` <= 1 ~ project1$`%rh` * 100,
                           project1$`%rh` > 1 ~ project1$`%rh`
                           )

# Project 1 - Missing data plot
# project1 %>% abbreviate_vars() %>% vis_miss()
## AQI Data

# AQI Data - Change marathon names and add year column
aqi_values_adj <- aqi_values %>%
  mutate(
    marathon = case_when(
      marathon == "Boston" ~ "B",
      marathon == "Chicago" ~ "C",
      marathon == "NYC" ~ "NY",
      marathon == "Twin Cities" ~ "TC",
      marathon == "Grandmas" ~ "D"
      ),
    year = year(date_local)
    )

# AQI Data - Missing data plot
# vis_miss(aqi_values)

# AQI Data - Filter data for simplicity
aqi_values_adj <- aqi_values_adj %>%
```

```
  filter(parameter_code %in% c(88101, 44201)) %>%
   ↪  select(-c("cbsa_code","state_code","county_code","site_number","date_local"))

# AQI Data - Create summary by year, race
aqi_summary_by_year <- aqi_values_adj %>%
  group_by(year, marathon, units_of_measure) %>%
  summarise(avg_arithmetic_mean=mean(arithmetic_mean, na.rm = TRUE))

aqi_summary_by_year <-
  spread(aqi_summary_by_year, units_of_measure, avg_arithmetic_mean) %>%
  arrange(year)

colnames(aqi_summary_by_year) <- c("Year", "Race", "PM2.5", "Ozone")

## Course Record
# Course Record - Missing data plot
# vis_miss(course_record)

# Course Record - Data Quality
course_record <- course_record %>% rename("Sex" = "Gender")
## Merging the Data
# Merge project 1 with course record
project1_CR <- project1 %>%
  left_join(course_record, by = c("Race", "Year", "Sex"))

# Calculate time
project1_CR$Time <- (project1_CR$CR * (1+project1_CR$`%CR`/100))/3600 %>%
  as.numeric()

# Merge project 1 + course record with AQI data
project1_CR_aqi <- project1_CR %>%
  left_join(aqi_summary_by_year, by = c("Race", "Year"))

# FIGURE 2: Correlation of full numeric data
library(corrplot)

project1_CR_aqi$Time <- project1_CR_aqi$Time %>% as.numeric()
data <- select_if(project1_CR_aqi,is.numeric) %>% na.omit() %>%
 ↪  abbreviate_vars()

M = cor(data)
```

```r
corrplot(M, method = 'number', type = "lower", diag = FALSE,
         number.cex = 0.5, cl.cex = 0.5,
         tl.cex = 0.65, tl.col = "black", tl.srt = 20)

###### AIM 1 ######

# FIGURE 3: Age vs Time Violin plot
project1_CR_aqi$AgeGroup <- cut(project1_CR_aqi$Age, breaks = seq(0, 100, by
 ↪  = 5))

ggplot(project1_CR_aqi, aes(x = AgeGroup, y = Time, fill = Sex)) +
    geom_violin(alpha = 0.6, scale = "width", position = position_dodge(0.8))
     ↪  +
    labs(x = "Age Group", y = "Course Time (Hours)") +
    theme_minimal() +
    theme(panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          axis.title = element_text(size = 8),
          axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
    scale_fill_manual(values = c("F" = "lightpink", "M" = "lightblue"))



# TABLE 4: Age & Time Summary by Sex
project1_CR_aqi %>% dplyr::select(Age, Sex, Time)%>%
  tbl_summary(by = Sex,
              type = list(where(is.numeric) ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({p25}, {p75})"))
               ↪  %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Age and Time Summary by Sex",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 10,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position"))

# Regression model - Time effects on Age
model_F <- lm(as.numeric(Time) ~ Age + I(Age^2),
              data = project1_CR_aqi[project1_CR_aqi$Sex == "F",])
model_M <- lm(as.numeric(Time) ~ Age + I(Age^2),
              data = project1_CR_aqi[project1_CR_aqi$Sex == "M",])
```

```r
# TABLE 5: Print results
summary_F <- summary(model_F)
summary_M <- summary(model_M)

coefficients_F <- as.data.frame(summary_F$coefficients)
coefficients_M <- as.data.frame(summary_M$coefficients)

coefficients <- cbind(coefficients_F, coefficients_M)
rownames(coefficients) <- c("Intercept", "Age", "Age Squared")

kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary",
      col.names = c("Estimate", "Std Error", "t value", "p value",
                    "Estimate", "Std Error", "t value", "p value")) %>%
  kableExtra::add_header_above(c(" ", "Female" = 4, "Male" = 4)) %>%
  kable_styling(font_size = 10, full_width = F, position = "center")
###### AIM 2 ######

# FIGURE 4: PM2.5 by Sex and Age
project1_CR_aqi$AgeGroup <- cut(project1_CR_aqi$Age, breaks = seq(0, 100, by
 ↪  = 10))

ggplot(project1_CR_aqi, aes(x = PM2.5, y = Time, color = Sex)) +
  # geom_point(alpha = 0.15, size = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "PM2.5 Levels", y = "Course Time (Hours)") +
  facet_wrap(~ AgeGroup, ncol = 5) +
  ylim(c(2.5,10)) +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_text(size = 8),
        axis.text = element_text(size = 8))

# FIGURE 5: Ozone by Sex and Age
ggplot(project1_CR_aqi, aes(x = Ozone, y = Time, color = Sex)) +
  # geom_point(alpha = 0.15, size = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Ozone", y = "Course Time (Hours)") +
  facet_wrap(~ AgeGroup, ncol = 5) +
  scale_x_continuous(breaks = c(0, 0.01, 0.03, 0.05)) +
  ylim(c(2.5,10)) +
```

```
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_text(size = 8),
        axis.text = element_text(size = 8))


project1_CR_aqi$AgeGroup <- ifelse(project1_CR_aqi$Age > 60, "60andAbove",
 ↪  "Under60") %>%
  factor(levels = c("Under60", "60andAbove"))

# Regression model, Time to AQI data
model <- lm(Time ~ Age * Sex + I(Age^2) +
              Age * PM2.5 + AgeGroup * PM2.5 + Sex * PM2.5 +
              Age * Ozone + AgeGroup * Ozone + Sex * Ozone,
            data = project1_CR_aqi)

backward_model <- step(model, direction = "backward", trace = 0)

# TABLE 6: Print results
summary <- summary(backward_model)
coefficients <- as.data.frame(summary$coefficients)
kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**t
       ↪  value**", "**p value**"))

###### AIM 3 ######

# Create summary table by race, year, WBGT, flag
new_data <- project1_CR_aqi %>% group_by(Race, Year, WBGT, Flag) %>%
 ↪  summarize(Avg_Time = mean(as.numeric(Time))) %>% na.omit()

# FIGURE 6: Plot average time by WBGT/Flag
ggplot(new_data, aes(x = Avg_Time, y = WBGT)) +
  geom_point(aes(color = Flag), alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "black") +
  labs(x = "Average Course Time",
       y = "WBGT",
       color = "Flag") +
  scale_color_manual(values = c("Green" = "green",
                                "Red" = "red",
```

```
                                  "White" = "grey",
                                  "Yellow" = "yellow"
                                )) +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_text(size = 8),
        legend.text = element_text(size = 8))

# Define age groups by quartiles
age_quartiles <- quantile(project1_CR_aqi$Age, probs = c(0, 0.25, 0.5, 0.75,
 ↪  1))
project1_CR_aqi <- project1_CR_aqi %>%
  mutate(AgeQuartile = cut(Age,
                           breaks = age_quartiles,
                           include.lowest = TRUE,
                           labels = c(paste("Under", age_quartiles[2]+1),
                                      paste0(age_quartiles[2]+1, "-",
                                       ↪  age_quartiles[3]),
                                      paste0(age_quartiles[3]+1, "-",
                                       ↪  age_quartiles[4]),
                                      paste0(age_quartiles[4]+1, "+"))
  ))


# FIGURE 7: Plot of %rh, SR, Wind, and WBGT by time
a <- ggplot(project1_CR_aqi, aes(x = `%rh`, y = as.numeric(Time), color =
 ↪  AgeQuartile)) +
    geom_smooth(method = "loess", se = FALSE) +
    geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
     ↪  "gray", linetype = "dashed") +  # Overall
    labs(x = "%rh", y = "Course Time (Hours)", title = "%rh") +
    theme_minimal() +
    theme(axis.title = element_text(size = 8),
          plot.title = element_text(size = 10),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.text = element_text(size = 8))

b <- ggplot(project1_CR_aqi, aes(x = SR, y = as.numeric(Time), color =
 ↪  AgeQuartile)) +
    geom_smooth(method = "loess", se = FALSE)  +
```

```
    geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
     ↪  "gray", linetype = "dashed") +  # Overall
    labs(x = "Solar Radiation", y = "", title = "Solar Radiation") +
    theme_minimal() +
    theme(axis.title = element_text(size = 8),
          plot.title = element_text(size = 10),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.text = element_text(size = 8))

c <- ggplot(project1_CR_aqi, aes(x = Wind, y = as.numeric(Time), color =
 ↪  AgeQuartile)) +
    geom_smooth(method = "loess", se = FALSE)  +
    geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
     ↪  "gray", linetype = "dashed") +  # Overall
    labs(x = "Wind Speed", y = "Course Time (Hours)", title = "Wind Speed") +
    theme_minimal() +
    theme(axis.title = element_text(size = 8),
          plot.title = element_text(size = 10),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.text = element_text(size = 8))

d <- ggplot(project1_CR_aqi, aes(x = WBGT, y = as.numeric(Time), color =
 ↪  AgeQuartile)) +
    geom_smooth(method = "loess", se = FALSE)  +
    geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
     ↪  "gray", linetype = "dashed") +  # Overall
    labs(x = "WBGT", y = "", title = "WBGT") +
    theme_minimal() +
    theme(axis.title = element_text(size = 8),
          plot.title = element_text(size = 10),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.text = element_text(size = 8))

grid.arrange(a, b, c, d, ncol = 2)

# Figure by sex
# a <- ggplot(project1_CR_aqi, aes(x = `%rh`, y = as.numeric(Time), color =
 ↪  Sex)) +
#     geom_smooth(method = "loess", se = FALSE) +
```

```
#     geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
 ↪  "gray", linetype = "dashed") +  # Overall
#     labs(x = "%rh", y = "Course Time (Hours)", title = "%rh") +
#     theme_minimal() +
#     theme(axis.title = element_text(size = 8),
#           plot.title = element_text(size = 10))
#
# b <- ggplot(project1_CR_aqi, aes(x = SR, y = as.numeric(Time), color =
 ↪  Sex)) +
#     geom_smooth(method = "loess", se = FALSE)  +
#     geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
 ↪  "gray", linetype = "dashed") +  # Overall
#     labs(x = "Solar Radiation", y = "", title = "Solar Radiation") +
#     theme_minimal() +
#     theme(axis.title = element_text(size = 8),
#           plot.title = element_text(size = 10))
#
# c <- ggplot(project1_CR_aqi, aes(x = Wind, y = as.numeric(Time), color =
 ↪  Sex)) +
#     geom_smooth(method = "loess", se = FALSE)  +
#     geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
 ↪  "gray", linetype = "dashed") +  # Overall
#     labs(x = "Wind Speed", y = "Course Time (Hours)", title = "Wind Speed")
 ↪  +
#     theme_minimal() +
#     theme(axis.title = element_text(size = 8),
#           plot.title = element_text(size = 10))
#
# d <- ggplot(project1_CR_aqi, aes(x = WBGT, y = as.numeric(Time), color =
 ↪  Sex)) +
#     geom_smooth(method = "loess", se = FALSE)  +
#     geom_smooth(aes(color = NULL), method = "loess", se = FALSE, color =
 ↪  "gray", linetype = "dashed") +  # Overall
#     labs(x = "WBGT", y = "", title = "WBGT") +
#     theme_minimal() +
#     theme(axis.title = element_text(size = 8),
#           plot.title = element_text(size = 10))
#
# grid.arrange(a, b, c, d, ncol = 2)

# Regression model - Course Time & Weather
model <- lm(Time ~ Age * (`%rh` + SR + Wind + WBGT + I(`%rh`^2) + I(SR^2) +
 ↪  I(WBGT^2)), data = project1_CR_aqi)
```

```r
# Backward selection model
final_model <- step(model, direction = "backward", trace = 0)
summary <- summary(final_model)

# TABLE 8: Print results
coefficients <- as.data.frame(summary$coefficients)

kable(coefficients, digits = 4,
      caption = "Linear Regression Model Summary: Backward Selection",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**t
       ↪  value**", "**Significance**"))
```