Impact of Environmental Conditions in Marathon Performance based on Gender and Age

Practical Data Analysis - Project 1: Exploratory Data Analysis

Aristofanis Rontogiannis

2024-9-29

Abstract

Purpose: This exploratory study examines the effects of environmental conditions, including temperature, humidity, solar radiation, wind, and air quality on marathon performance. Specifically, it explores how these conditions influence performance across different ages and genders, and identifies the weather parameters with the most significant impact on performance outcomes.

Methods: Data from five major marathons over a 15-20 year period, covering ages 14 to 85, were analyzed. Environmental conditions such as temperature (°C), humidity (%), wind speed (km/hr), WBGT (°C) and others were recorded. Performance was measured based on finish times, and exploratory analyses were conducted to identify the relationship between weather conditions and race performance.

Results: Exploratory analyses indicate that Solar Radiation and some pollutants in the air have the greatest influence on marathon performance. Older runners and children (mostly female athletes) showed greater sensitivity to environmental conditions, leading to slower finish times compared to middle aged male participants.

Conclusions: Environmental factors, particularly air quality (some pollutants in the air) and solar radiation, play a critical role in marathon performance. Those findings can help inform runners and organizers about the potential impacts of weather on race outcomes and guide race-day decisions.

Introduction

This project is a collaboration with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. The research aims to explore the relationship between environmental conditions and endurance exercise performance, specifically within the context of marathon races. Endurance performance is known to decline with increasing environmental temperatures (B. R. Ely et al. (2010)), a trend that becomes more pronounced during longer-distance events, such as the marathon (M. R. Ely et al. (2007)). This decline is further exacerbated among older adults, who experience thermoregulatory challenges that hinder their ability to dissipate heat efficiently (Kenney and Munce (2003)).

Moreover, differences between men and women in endurance performance are well-documented (Besson et al. (2022)), as well as in physiological processes related to thermoregulation (Yanovich, Ketko, and Charkoudian (2020)). These factors collectively contribute to variations in performance outcomes during marathons under different environmental conditions.

This dataset includes top performances from five major marathons spanning 15 to 20 years. It covers athletes aged 14 to 85, detailing not only performance but also comprehensive environmental data. The primary objective of this research is to analyze how variables like temperature, humidity, solar radiation, wind, and WBGT (Wet-Bulb Globe Temperature) impact marathon performance across different age groups and between genders.

The project focuses on the effects of increasing age on marathon performance in both men and women, on how environmental conditions influence marathon performance with a focus on potential differences across age and gender and which weather parameters have the most significant impact on marathon performance.

Data Analysis

On this project, we have four datasets which include marathon performance data, air quality index (AQI) values, marathon dates, and course records. We inspect the dimensions and structure of each dataset and identify any missing values. More specificaly we create the percentage of missingness on each variable of each dataset. This preliminary examination is crucial for understanding the completeness of our data and assessing how missing values might impact our analysis. Below, you can see the missingness table for the first dataset. Note that this table is based mostly on raw data. We observe that there is a pattern in the missingness. Eight of the sixteen variables on the first dataset have the same percentage of missingness. We will discuss this pattern later in our analysis.

Table 1: Percentage of Missing Values by Variable

Variable	Missing_Percentage
Td..C	4.245936
Tw..C	4.245936
X.rh	4.245936
Tg..C	4.245936
SR.W.m2	4.245936
DP	4.245936
Wind	4.245936
WBGT	4.245936
Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.	0.000000
Year	0.000000
Sex..0.F..1.M.	0.000000
Flag	0.000000
Age..yr.	0.000000
X.CR	0.000000

In order to examine the effects of age in marathon performance, we categorize the age of marathon participants into defined groups. This means we create a factor variable named "age_group" that contains nine age groups. The age groups are as follows: "14 and under", "15-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", and "80 and over." I think this is helpful as this stratification allows for more nuanced analysis of performance across different age ranges. Note that in our analysis below we will use both the factor variable "age_group" and the continuous one.

At this point, we want to find the effects of increasing age on marathon performance in men and women. So, the next step involves merging two datasets, more specifically data1 and data4. These two datasets share common variables, more specifically variables related to gender and race (marathon). We first standardize the variable names and structures to facilitate this merge. This allows us to create a comprehensive dataset (data14) that includes marathon performance metrics alongside demographic and some environmental data.

We create the summary table of the merged dataset tha shows the environmental factors based on the Race. For each of those variables we calculate the mean, sd, min and max if they are continuous, and their number and percentage if they are factor. The 0, 1, 2, 3 and 4 are the Boston marathon, the Chicago marathon, the New York City marathon, the Twin Cities marathon and the Grandmas marathon accordingly. Gender values are 0, 1 for Male and Female. Note that this table is based mostly on raw data and does not include environmental factors related to air quality. Later in our analysis, we will combine all given datasets and another summary table is created.

Table 2: Summary of Environmental Factors by Marathon, N = 11564

Characteristic	**0** N = 2,088	**1** N = 2,553	**2** N = 2,930	**3** N = 1,993	**4** N = 2,000
Sex/Gender of the runners					
0	984 (47%)	1,210 (47%)	1,402 (48%)	922 (46%)	934 (47%)
1	1,104 (53%)	1,343 (53%)	1,528 (52%)	1,071 (54%)	1,066 (53%)
WBGT (C)					
Mean (SD)	11.32 (4.53)	12.12 (5.76)	10.74 (4.91)	13.20 (5.35)	18.65 (3.22)
Min, Max	6.55, 23.20	1.35, 24.74	3.65, 18.87	6.51, 24.22	14.03, 25.13
WBGT flag					
	0 (0%)	126 (4.9%)	131 (4.5%)	118 (5.9%)	116 (5.8%)
Green	810 (39%)	1,459 (57%)	901 (31%)	834 (42%)	702 (35%)
Red	123 (5.9%)	116 (4.5%)	0 (0%)	116 (5.8%)	237 (12%)
White	1,040 (50%)	732 (29%)	1,394 (48%)	587 (29%)	0 (0%)
Yellow	115 (5.5%)	120 (4.7%)	504 (17%)	338 (17%)	945 (47%)
Age in years					
Mean (SD)	46.96 (17.26)	45.82 (17.89)	49.57 (18.76)	44.72 (17.44)	44.03 (17.51)
Min, Max	18.00, 87.00	14.00, 85.00	18.00, 91.00	14.00, 85.00	14.00, 86.00
Wind speed (Km/hr)					
Mean (SD)	11.99 (4.47)	8.21 (3.19)	11.22 (4.55)	8.80 (3.20)	9.16 (2.87)
Min, Max	4.75, 21.75	3.00, 16.25	0.00, 20.00	3.67, 15.67	3.78, 14.00
Dry bulb temperature (C)					
Mean (SD)	11.64 (5.89)	12.42 (6.05)	11.73 (4.67)	13.14 (5.52)	18.86 (3.32)
Min, Max	5.33, 28.14	2.00, 25.67	5.25, 20.00	7.00, 24.67	13.00, 24.62
Wet bulb temperature (C)					
Mean (SD)	7.59 (3.81)	8.54 (5.68)	7.57 (4.98)	9.85 (5.41)	14.91 (2.45)
Min, Max	2.50, 17.54	-1.27, 21.49	0.61, 17.00	1.98, 21.60	9.99, 19.68
Percent relative humidity					
Mean (SD)	36.11 (34.18)	60.45 (10.49)	26.89 (30.44)	41.78 (34.23)	49.34 (34.28)
Min, Max	0.28, 98.25	43.00, 85.00	0.31, 98.33	0.37, 89.13	0.42, 89.67
Black globe temperature (C)					
Mean (SD)	24.22 (8.38)	24.52 (6.30)	21.36 (5.93)	24.94 (6.52)	31.63 (7.86)
Min, Max	9.51, 42.39	10.17, 35.66	11.37, 34.78	12.67, 35.38	13.93, 44.45
Dew Point (C)					
Mean (SD)	3.32 (4.47)	4.65 (6.86)	2.74 (6.99)	5.96 (7.25)	12.43 (3.17)
Min, Max	-4.40, 13.51	-7.00, 19.67	-7.25, 16.20	-7.43, 20.33	4.00, 18.00
SR.W.m2					
Mean (SD)	649.80 (186.92)	460.48 (94.56)	401.15 (130.90)	435.89 (138.93)	676.81 (190.51)
Min, Max	147.15, 852.69	252.80, 608.47	142.73, 573.41	141.37, 630.15	289.45, 909.47

[1] n (%)

At this point, we want to address this pattern of missingness in our dataset. The variable “Flag” is an indicator of Wet Bulb Globe Temperature and risk of heat illness. With that saying, the “Flag” variable can be “White”, “Green”, “Yellow”, “Red” and “Black” based on the amount of heat and high temperature. We noticed that our dataset contains data for all flags except “Black”. There are some data in the “Flag” variable tho, that are categorized as ““. This might be the case because when the”Flag” is “Black” the race does not take place. What I am trying to say is the pattern of missingness is the result of this scenario. So at this point I find it reasonable to delete those data as those races (marathons) did not take place.

To ensure the integrity of our analysis, we again evaluate the percentage of missing values in the merged dataset. This step is essential for identifying any potential gaps in the data that could affect our results. As we have assumed, now there are no missing values in our dataset.

In the merged dataset, there is a variable named "CR" or course record. We have to convert the course record (CR) from a character string format into seconds (numeric), which allows straightforward comparisons. Following this transformation, we create a new variable, performanceOpp, which indicates a runner's performance relative to the course record. In this context, a higher value for performanceOpp represents a slower time, thereby reflecting the relationship between performance and the current course record.

Analysis of Performance Based on Gender and Age Groups

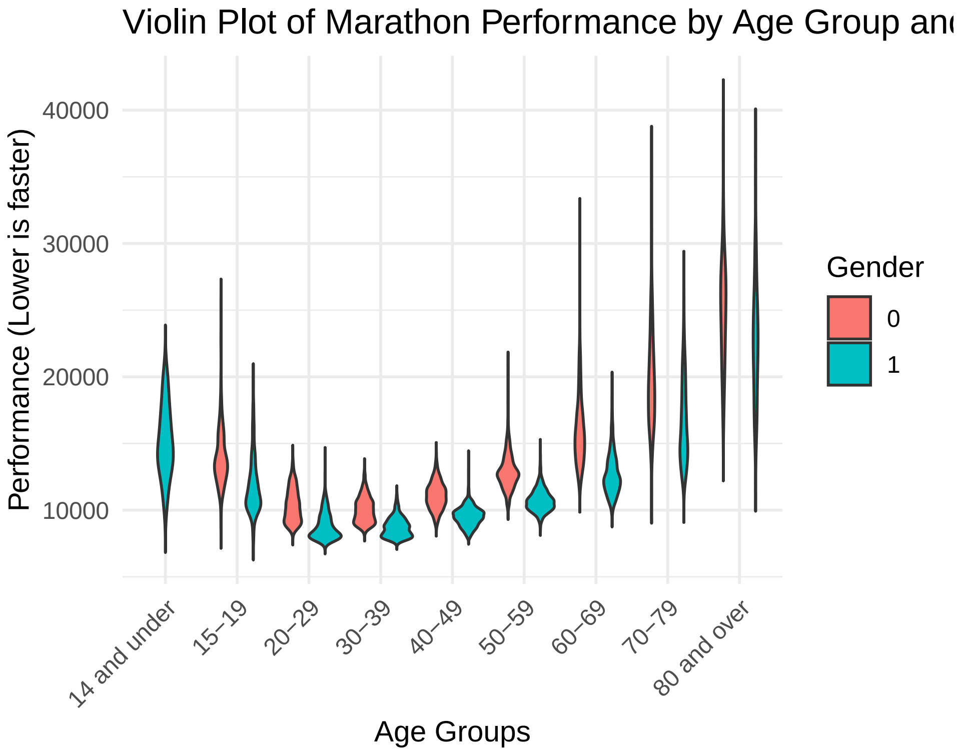
Using the new dataset, we calculate summary statistics to analyze mean and median performances across different age groups and genders. The results are printed in a summary table, which provides insights into the performance trends. After that, we sort those data based on the mean performance variable. First of all, there are only male athletes aged less than 14 years old (n=23 boys). We can clearly notice that males (Gender=1) aged 30-39 and 20-29 are the ones achieving the best performance, followed by males aged 40-49 years old. The best female runners (Gender=0), based on their performance, belong to the age group of 30-39 years old. Those results are pretty reasonable for marathon running events.

Table 3: Summary of Mean and Median Performance by Age Group and Gender

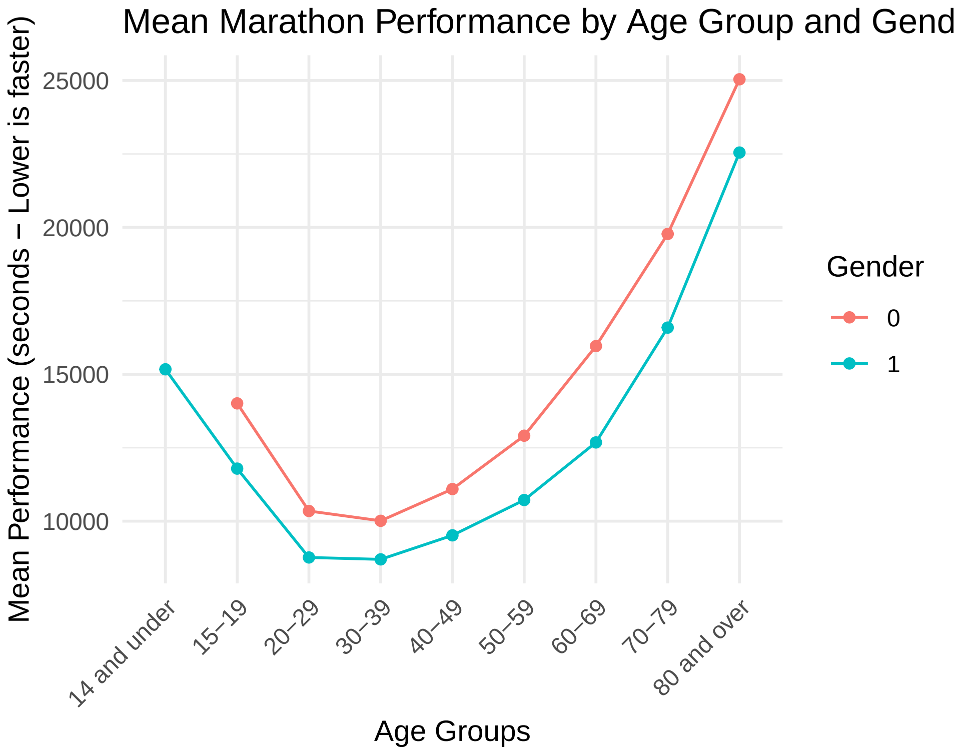
age_group	Gender	mean_performance	median_performance	count
30-39	1	8699.052	8633.5	920
20-29	1	8764.058	8455.0	919
40-49	1	9517.246	9525.5	920
30-39	0	10011.571	9917.0	920
20-29	0	10345.316	10192.0	920
50-59	1	10719.133	10632.0	919
40-49	0	11092.868	11062.5	920
15-19	1	11788.794	11246.5	316
60-69	1	12680.538	12397.5	916
50-59	0	12908.937	12749.0	919
15-19	0	14009.170	13586.5	294
14 and under	1	15169.348	14791.0	23
60-69	0	15959.739	15570.5	830
70-79	1	16589.159	15952.5	742
70-79	0	19776.769	19307.0	368
80 and over	1	22547.372	22488.0	180
80 and over	0	25041.638	25555.0	47

In order to evaluate our results and examine the effects of increasing age on marathon performance in men and women, we have to visualize the data. We create various plots, including boxplots, line plots, and violin plots, to illustrate the performance distributions by age group and gender. Finally, we include histograms to explore the distribution of performance across age groups, something that provides an intuitive understanding of the performance landscape within our dataset. These visualizations facilitate a deeper understanding of how marathon performance varies with age and gender.

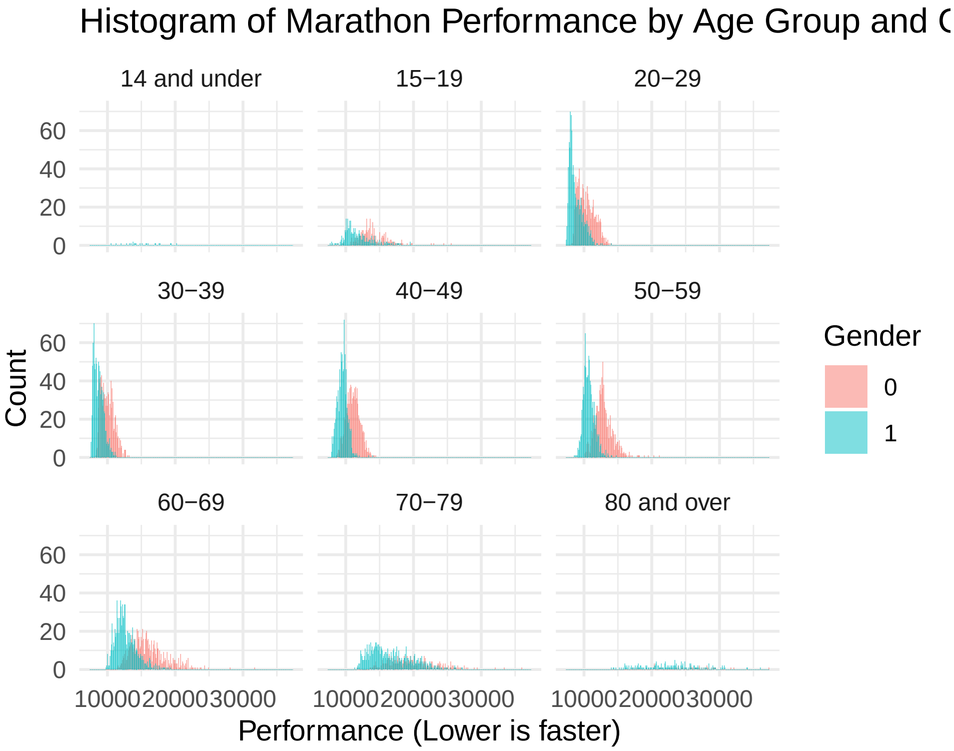
In the violin plot below, we can clearly see that the age groups of 20-29, 30-39 and 40-49 years old achieved a better performance on both male and female athletes, while people aged more than 80 years old are the slowest ones, something that is completely reasonable. Those age groups also have less variability and less outliers. In contrast, people that are younger or older than that seem to correspond to data with more variability meaning that some runners did much longer or shorter times than others. Those outliers can affect the results of an analysis, such as skewing means, impacting regression results, or influencing correlation coefficients. This specific project does not aim in regression analysis or modeling so we will not delete or trying to use them.



Furthermore, we create a line plot to visualize the mean performance by age group and gender. The interpretation of this plot is the same as of the above plots and tables. We can clearly see that the line is a "u-shape" line, meaning that children, teenagers and older athletes tend to run slower than the others, with male athletes running generally faster than female ones in all age groups.



We also create histograms to explore the distribution of performance across age groups and gender. The first thing we notice is that the majority of people that take place in these marathons are aged 20 to 69 years old. Also, in those visualizations, the more to the left the plot is, the better the performance. So we can clearly see that the age groups of 20-29, 30-39 and 40-49 years old achieved a better performance on both male and female athletes.

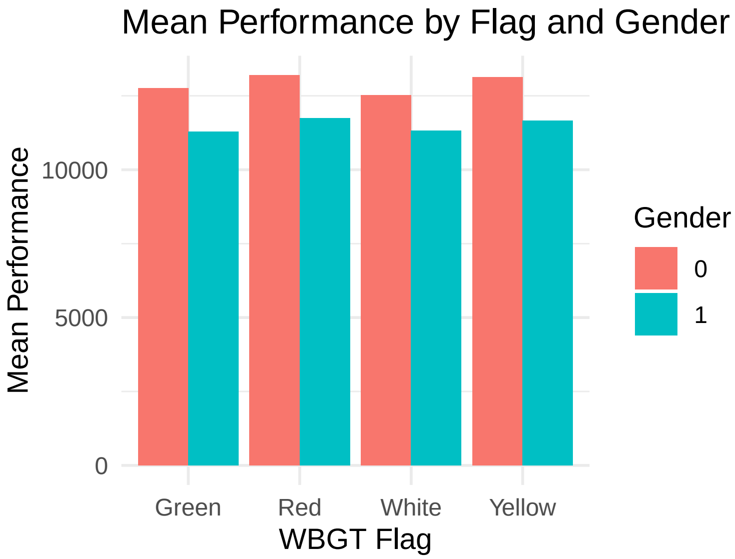


Analysis of Performance Based on Flag and Gender

In this section, we aim to investigate how the Flag (based on WBGT), which indicates varying levels of temperature, affects marathon performance across different genders. To facilitate our analysis, the Flag and Gender variables in the dataset have been converted into factors for better handling.

The summary statistics were calculated to examine the average performance (performanceOpp, where lower values indicate better performance) of marathon runners based on the WBGT-flag status and gender. From the summary statistics table, it is observed that male runners consistently achieve faster times than female runners across all temperature conditions. This suggests that gender plays a significant role in performance outcomes under varying temperature conditions.

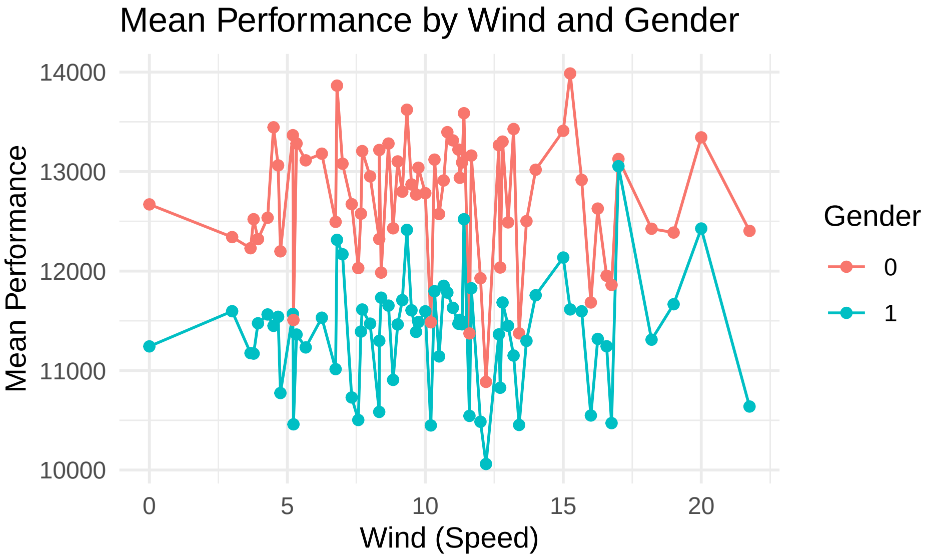
The bar plot below provides a clear comparison of mean performance across different WBGT flag conditions, segmented by gender. The results highlight noticeable differences in performance levels, emphasizing that as the temperature conditions indicated by the WBGT flag become more strenuous, the performance gap between male and female runners persists. Another important thing is that the temperature does not seem to affect the performance. There is only a slight decrease in the performance when the temperature conditions are worst. This might happen because at this point we used the variable WBGT which is the weighted average of dry bulb, wet bulb, and globe temperature and we do not know the relationship between each of those variables and performance yet.



Analysis of Performance Based on Wind and Gender

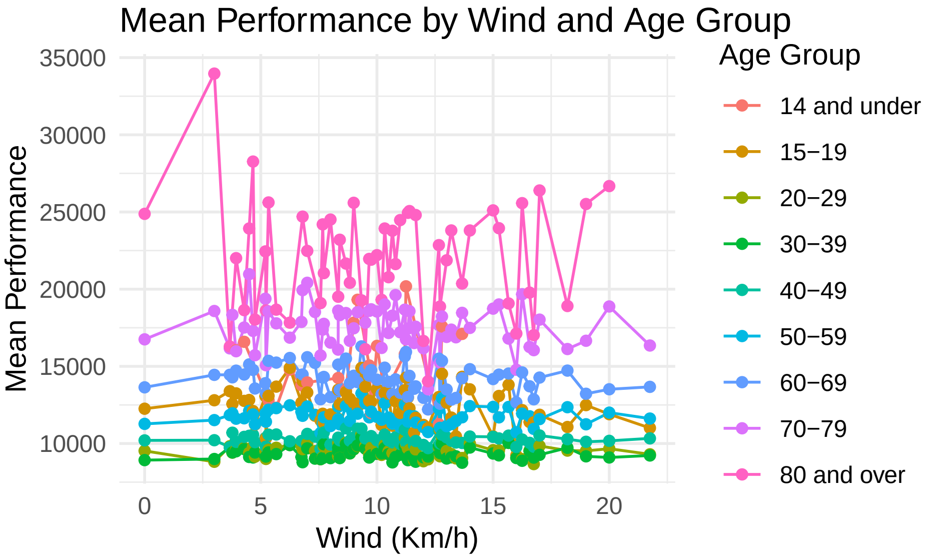
Now we are interested in finding any relationship between Wind and Performance. For both genders the correlation is pretty weak. So both male and female performance is not affected from wind using our data. An explanation to this is that the range of wind in our data is from 0 to 21.75 km/h, so not so strong.

The plot below shows the average performance based on wind speed and gender. As observed in the plot, there is no clear pattern or strong correlation between wind speed and performance, suggesting that wind speed might not have a significant impact on marathon times in this dataset. However, for young and old athletes (<14yo and >80yo) we can see a relationship. It is evident that, on average, men tend to perform better than women across all wind conditions. This aligns with general observations of gender differences in endurance performance, but the impact of wind speed itself does not appear to be substantial.



Analysis of Performance Based on Wind and Age Groups

The plot below shows the average performance based on wind speed and age group. As observed in the plot, there is no clear pattern or strong correlation between wind speed and age group, suggesting that wind speed might not have a significant impact on marathon times in this dataset. However, it is evident that, on average, people aged 20-49 years old tend to run faster across all wind conditions. This aligns with general observations of age group differences in endurance performance, but the impact of wind speed itself does not appear to be substantial.



We can see that none of the temperature values or Wind affects the performance. We also calculated the correlation of other variables like Td, Tw, Rh and SR with performance and it is still weak. This might happen because some of those environmental variables relate to each other.

Analysis of the Environmental Variables that have the biggest impact in Performance

After all that said, we want to incorporate the air quality into our analysis to examine if it affects the performance. The variable aqi contains a lot of missing values and it may not be helpful to use it in our analysis.

We create a new dataset that combines our old dataset with a new one. The new dataset contains local pollutants measured in the aqi like PM2.5, Sulfur Dioxide, Nitrogen Dioxide and Ozone that we extracted with the r packagce RAQSAPI.

At this point we will create a summary table for this dataset. This is the merged dataset containing the most important environmental variables that can have an impact on performance. for temperature, we using the variable WBGT which is Weighted average of dry bulb, wet bulb, and globe temperature. For each of those variables we calculate the mean, sd, min and max if they are continuous, and their number and percentage if they are factor. The 0, 1, 2, 3 and 4 are the Boston marathon, the Chicago marathon, the New York City marathon, the Twin Cities marathon and the Grandmas marathon accordingly. Gender values are 0, 1 for Male and Female.

We can see that the variable PM2.5 has a lot of missing values (70% missing) so it will be better if we delete it.

As environmental variables we will use the WBGT for temperature (it is Weighted average of dry bulb, wet bulb, and globe temperature - the other temperature variables are correlated to each other), the Wind, the humidity, the solar radiation and the four pollutants (Ozone, Nitrogen Dioxide, Sulfur Dioxide and PM2.5)

Table 4: Percentage of Missing Values by Variable

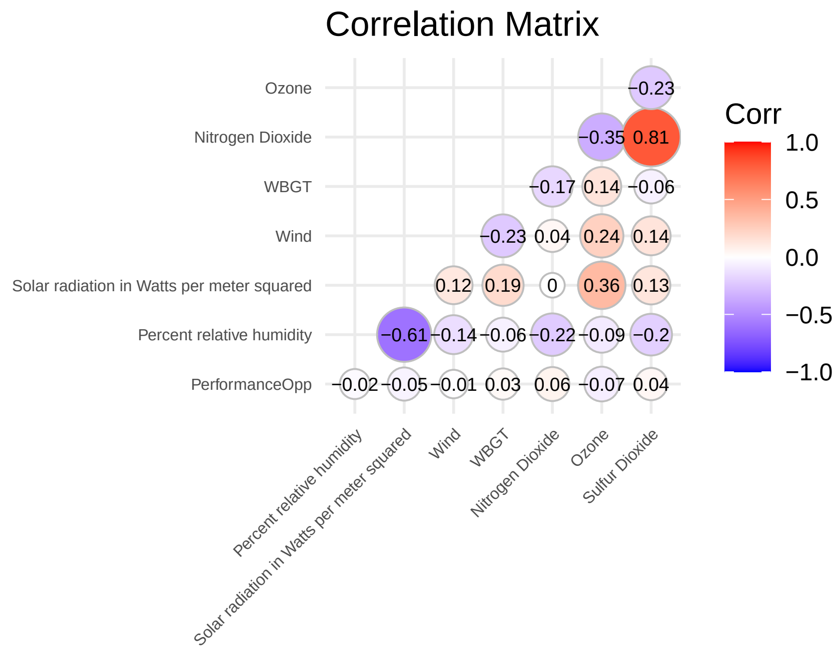
Variable	Missing_Percentage
Nitrogen Dioxide	20.710827
Sulfur Dioxide	19.802836
Ozone	7.497406
marathon/race	4.245936
Year	4.245936
Sex	4.245936
Flag	4.245936
Age	4.245936
Percent relative humidity	4.245936
Solar radiation in Watts per meter squared	4.245936
Wind	4.245936
WBGT	4.245936
Gender	4.245936
PerformanceOpp	4.245936
age_group	4.245936

Table 5: Summary of Environmental Factors by Marathon, N = 11073

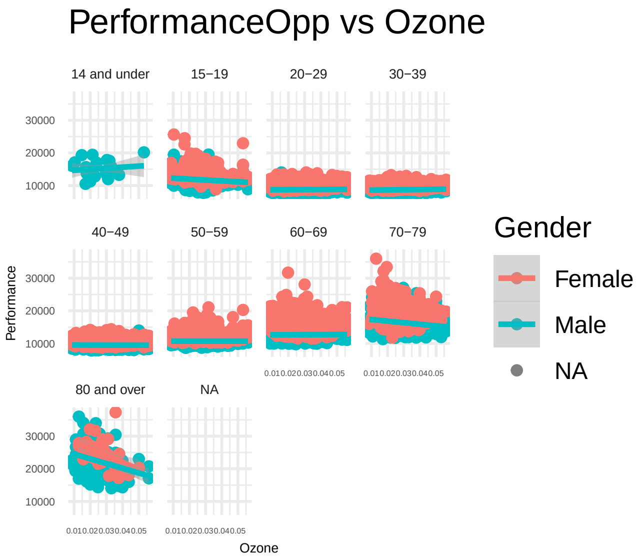
Characteristic	**0**	**1**	**2**	**3**	**4**
	N = 2,088	N = 2,427	N = 2,799	N = 1,875	N = 1,884
Gender of the runners					
0	984 (47%)	1,150 (47%)	1,337 (48%)	867 (46%)	880 (47%)
1	1,104 (53%)	1,277 (53%)	1,462 (52%)	1,008 (54%)	1,004 (53%)
WBGT (C)					
Mean (SD)	11.32 (4.53)	12.12 (5.76)	10.74 (4.91)	13.20 (5.35)	18.65 (3.22)
Min, Max	6.55, 23.20	1.35, 24.74	3.65, 18.87	6.51, 24.22	14.03, 25.13
Age in years					
Mean (SD)	46.96 (17.26)	45.75 (17.85)	49.53 (18.75)	44.72 (17.44)	44.02 (17.53)
Min, Max	18.00, 87.00	14.00, 85.00	18.00, 91.00	14.00, 85.00	14.00, 85.00
WBGT flag					
White	1,040 (50%)	732 (30%)	1,394 (50%)	587 (31%)	0 (0%)
Green	810 (39%)	1,459 (60%)	901 (32%)	834 (44%)	702 (37%)
Yellow	115 (5.5%)	120 (4.9%)	504 (18%)	338 (18%)	945 (50%)
Red	123 (5.9%)	116 (4.8%)	0 (0%)	116 (6.2%)	237 (13%)
Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Wind speed (Km/hr)					
Mean (SD)	11.99 (4.47)	8.21 (3.19)	11.22 (4.55)	8.80 (3.20)	9.16 (2.87)
Min, Max	4.75, 21.75	3.00, 16.25	0.00, 20.00	3.67, 15.67	3.78, 14.00
Percent relative humidity					
Mean (SD)	36.11 (34.18)	60.45 (10.49)	26.89 (30.44)	41.78 (34.23)	49.34 (34.28)
Min, Max	0.28, 98.25	43.00, 85.00	0.31, 98.33	0.37, 89.13	0.42, 89.67
Solar radiation (W/m^2)					
Mean (SD)	649.80 (186.92)	460.48 (94.56)	401.15 (130.90)	435.89 (138.93)	676.81 (190.51)
Min, Max	147.15, 852.69	252.80, 608.47	142.73, 573.41	141.37, 630.15	289.45, 909.47
Ozone pollutant					
Mean (SD)	0.04 (0.01)	0.02 (0.00)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)
Min, Max	0.03, 0.06	0.01, 0.03	0.01, 0.04	0.01, 0.05	0.02, 0.05
SO2 pollutant					
Mean (SD)	2.25 (1.39)	3.71 (2.84)	3.55 (2.03)	0.71 (0.68)	1.05 (1.22)
Min, Max	0.33, 4.94	0.15, 10.36	1.10, 7.87	0.00, 2.95	0.04, 3.14
NO2 pollutant					
Mean (SD)	9.11 (3.13)	15.89 (7.44)	18.55 (5.62)	6.98 (2.73)	0.93 (1.09)
Min, Max	4.04, 14.68	3.80, 28.94	8.57, 28.37	3.74, 13.68	0.00, 2.46
Seconds off record					
Mean (SD)	11,260.26 (2,754.52)	11,964.40 (3,711.66)	12,555.57 (4,559.98)	12,056.78 (3,071.59)	12,229.25 (3,371.18)
Min, Max	7,382.00, 25,383.00	7,425.00, 33,965.00	7,662.00, 37,312.00	7,731.00, 22,607.00	7,746.00, 28,082.00

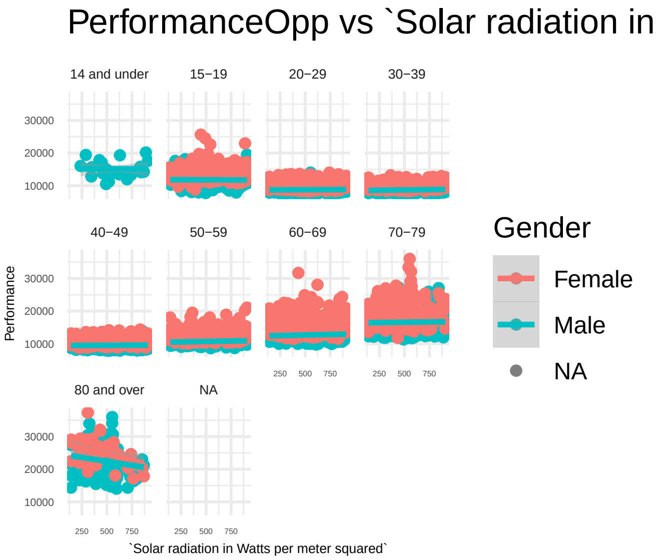
[1] n (%)

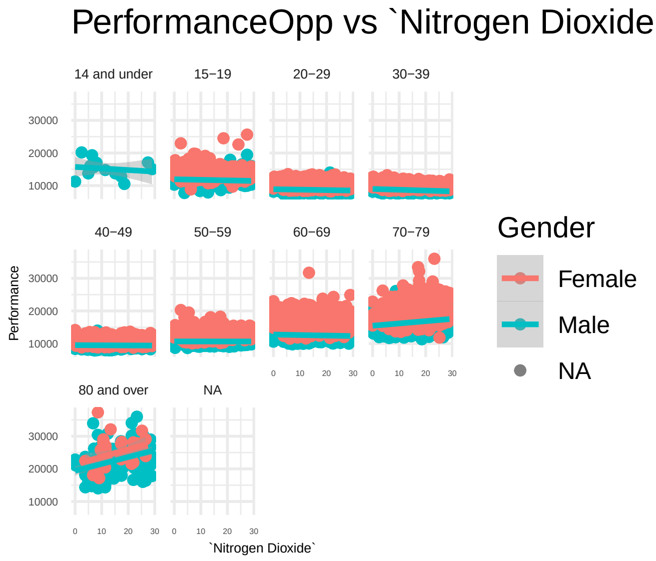
At this point we want to examine which of the seven environmental variables affect the performance. First of all, we have to examine if they are correlated with the variable performance at all. We observe that all seven environmental variables have weak correlation to performance. Based on the correlations, the variable with the most impact on performance is Ozone, followed by Nitrogen Dioxide, Solar Radiation and Sulfur Dioxide. . However, there are some variables that are correlated with each other. For example, Sulfur with Nitrogen dioxide, and relative humidity with solar radiation.

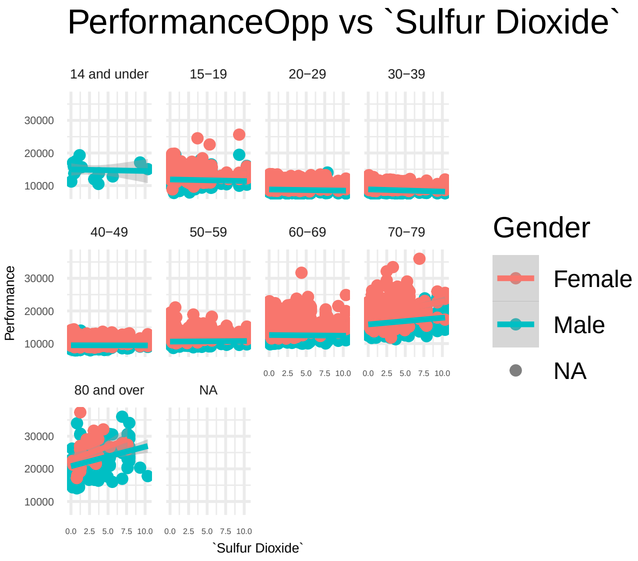


Those four plots below show that Ozone, Nitrogen Dioxide, Sulfur Dioxide and Solar radiation affect performance. Those environmental conditions tho, mostly affect athletes that are either too young (less that 14 years old) or old. Furthermore, female athletes seem to be more susceptible to those environmental conditions than male.









A more sophisticated approach is to perform linear regression. This will help us understand the relationship between performance and the environmental variables, while controlling for the effects of other variables. Since the two dioxides are strongly correlated, we can use the one with the strongest correlation to the performance (Nitrogen Dioxide). Similarly, we will use Solar radiation and not the humidity. This approach can also help us examine which environmental condition has a strongest impact on Performance.

After running some linear regression models, we can conclude that linearity is not suitable for those data. Although,it is obvious that Age and Gender have huge impact in the Performance, nothing can be concluded for the environmental factors as it seems that their relationship with Performance might not be linear. At this point someone can proceed by doing Splines or Generalized Additive Models or something else that can examine non-linear relationships. We opted not to include linear regressions or anything else on the report of this project as it focuses more in exploratory analysis.

Conclusions and Limitations

The findings from this exploratory analysis suggest that air quality and solar radiation have an impact on marathon performance, particularly for older and younger participants. Male middle-aged runners performed better on average, while wind and humidity had a less pronounced effect on their performance. Additionally, the decline in performance was more substantial in older age groups, indicating that thermoregulatory challenges increase with age. These results provide valuable insights for race organizers and athletes, emphasizing the need to consider environmental conditions when planning training and races.

This study has several limitations that should be considered when interpreting the results. First, a portion of the data were missing, something that may have affected the accuracy of the analysis. Additionally, the dataset only included a few major marathons, limiting the generalizability of the findings to other races or geographical regions. Furthermore, individual factors such as fitness level, nutrition, and race strategy, which also likely influence performance, were not included in the analysis. Future research should consider incorporating these variables and collecting a more comprehensive dataset over a wider range of conditions.

Data Privacy and Code Availability

The analysis dataset was obtained by Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. The replication code can be found at https://github.com/AristofanisR/Practical_Data_Analysis_Project1

References

Besson, Thibault, Robin Macchi, Jeremy Rossi, Cédric YM Morio, Yoko Kunimasa, Caroline Nicol, Fabrice Vercruyssen, and Guillaume Y Millet. 2022. "Sex Differences in Endurance Running." *Sports Medicine* 52 (6): 1235–57.

Ely, Brett R, Samuel N Cheuvront, Robert W Kenefick, and Michael N Sawka. 2010. "Aerobic Performance Is Degraded, Despite Modest Hyperthermia, in Hot Environments." *Med Sci Sports Exerc* 42 (1): 135–41.

Ely, Matthew R, Samuel N Cheuvront, William O Roberts, and Scott J Montain. 2007. "Impact of Weather on Marathon-Running Performance." *Medicine and Science in Sports and Exercise* 39 (3): 487–93.

Kenney, W Larry, and Thayne A Munce. 2003. "Invited Review: Aging and Human Temperature Regulation." *Journal of Applied Physiology* 95 (6): 2598–2603.

Yanovich, R, I Ketko, and N Charkoudian. 2020. "Sex Differences in Human Thermoregulation: Relevance for 2020 and Beyond." *Physiology* 35 (3): 177–84.

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      error = FALSE)

# install.packages("dplyr")
#install.packages("glue")
#install.packages("patchwork")
#install.packages("lubridate")
#install.packages("ggplot2")
#install.packages("reshape2")
#install.packages("GGally")
#install.packages("ggcorrplot")
#install.packages("gtsummary")
#install.packages("summarytools")
library(summarytools)
library(ggplot2)
library(knitr)
library(kableExtra)
library(GGally)
library(patchwork)
library(dplyr)
library(reshape2)
library(tidyr)
library(lubridate)
library(gtsummary)
library(gt)
library(ggcorrplot)
#First of all lets read all datasets
data1<-read.csv("project1.csv")
data2<-read.csv("aqi_values.csv")
data3<-read.csv("marathon_dates.csv")
data4<-read.csv("C:/Users/arontog1/Downloads/course_record (1).csv")
# dim(data1)
# dim(data2)
# dim(data3)
# dim(data4)

#View(data1)
#dim(data1)
#sum(is.na(data1))
#str(data1)
#colnames(data1)

#first, lets find the percentage of missing values on each column on dataset 1
#  Calculate the percentage of missing values for each column
missingness_table <- data1 %>%
  summarise(across(everything(), ~ sum(is.na(.)) / n() * 100)) %>%
  pivot_longer(everything(), names_to = "Variable",
               values_to = "Missing_Percentage") %>%
  arrange(desc(Missing_Percentage))
```

```
# Create a formatted table with kableExtra
kable(missingness_table, format = "latex", booktabs = TRUE,
      caption = "Percentage of Missing Values by Variable") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down"),
                position = "center",
                full_width = FALSE) %>%
  column_spec(1, bold = TRUE) %>%  # Make the "Variable" column bold
  row_spec(0, bold = TRUE)         # Make header row bold
#lets find the percentage of missing values on each column on dataset 2
mis_perc<-colSums(is.na(data2)) / nrow(data2) * 100
mis_table <- data.frame(Missing_Percentage = mis_perc)
#sort(mis_perc)

#lets find the percentage of missing values on each column on dataset 3
mis_perc<-colSums(is.na(data3)) / nrow(data3) * 100
mis_table <- data.frame(Missing_Percentage = mis_perc)
#sort(mis_perc)

#lets find the percentage of missing values on each column on dataset 4
mis_perc<-colSums(is.na(data4)) / nrow(data4) * 100
mis_table <- data.frame(Missing_Percentage = mis_perc)
#sort(mis_perc)

#First of all, I think it would be helpful to create age groups.

data1 <- data1 %>%
  mutate(age_group = cut(
    Age..yr.,
    breaks = c(-Inf, 14, 19, 29, 39, 49, 59, 69, 79, Inf),
    labels = c(
      "14 and under",
      "15-19",
      "20-29",
      "30-39",
      "40-49",
      "50-59",
      "60-69",
      "70-79",
      "80 and over"
    ),
    right = TRUE
  ))

#I want to combine data1 and data4 as they share the variable sex and race
#First i wanna rename "Sex..0.F..1.M." to "Gender" and the
#"Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D." to "Race"
names(data1)[names(data1) == "Sex..0.F..1.M."] <- "Gender"
names(data1)[names(data1) == "Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D."] <- "Race"

#now I have to transform the data4$Race to take values 0, 1, 2, 3, 4
#so they can match the data1$Race
data4 <- data4 %>%
  mutate(Race = recode(
    Race,
    "B" = 0,
    "C" = 1,
    "NY" = 2,
    "TC" = 3,
    "D" = 4
  ))

#now I have to transform the data4$Gender to take values 0, 1 so they can match
#the data1$Gender
data4 <- data4 %>%
  mutate(Gender = recode(Gender, "M" = 1, "F" = 0, ))
#So now I  will combine data1 and data4 as they share the variable sex and race
data14 <- left_join(data1, data4, by = c("Race", "Gender", "Year"))
#View(data14)
#str(data14)
data14$Age..yr.<-as.numeric(data14$Age..yr.)
data14$Race<-as.factor(data14$Race)
data14$Gender<-as.factor(data14$Gender)


data14 %>%
  dplyr::select(Race,
                Gender,
                WBGT,
                Flag,
                Age..yr.,
                Wind,
                Td..C,
                Tw..C,
                X.rh,
                Tg..C,
                DP,
                SR.W.m2) %>%
  distinct() %>%
  tbl_summary(
    statistic = list(
      all_continuous() ~ c("{mean} ({sd})", "{min}, {max}"),
      all_categorical() ~ "{n} ({p}%)"
    ),
    by = Race,
    # Group by marathon
    digits = all_continuous() ~ 2,
    missing = "no",
    type = list(
      WBGT ~ "continuous2",
      Gender ~ "categorical",
      Flag ~ "categorical",
      X.rh ~ "continuous2",
      Wind ~ "continuous2",
      Age..yr. ~ "continuous2",
      Tw..C ~ "continuous2",
      Tg..C ~ "continuous2",
```

```r
      Td..C ~ "continuous2",
      DP ~ "continuous2",
      SR.W.m2  ~ "continuous2"
    ),
    label = list(
      WBGT = "WBGT (C)",
      Gender = "Sex/Gender of the runners",
      Flag = "WBGT flag",
      Age..yr. = "Age in years",
      X.rh = "Percent relative humidity",
      Wind = "Wind speed (Km/hr)",
      SR = "Solar radiation (W/m^2)",
      Tw..C = "Wet bulb temperature (C) ",
      Tg..C = "Black globe temperature (C) ",
      Td..C = "Dry bulb temperature (C)",
      DP = "Dew Point (C)"
    )
  ) %>%
  modify_caption(caption = "Summary of Environmental Factors by Marathon,
                 N = {N}") %>%
  as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = "",
    format = "latex"
  ) %>%
  kableExtra::kable_styling(
    position = "center",
    latex_options = c("striped", "repeat_header"),
    stripe_color = "gray!15",
    font_size = 8
  )
#lets find the percentage of missing values on each column on dataset 14
mis_perc <- colSums(is.na(data14)) / nrow(data14) * 100
mis_table <- data.frame(Missing_Percentage = mis_perc)
#sort(mis_perc)
data14$Flag <- as.factor(data14$Flag)
#levels(data14$Flag)
#sum(data14$Flag == "")
#For some reason "Black" flag is presented as "" in the dataset. This is because
#if the Flag is Black the race is not taking place. So I will delete the races
#with (data14$Flag) == ""
data14 <- data14[data14$Flag != "", ]
#lets find the percentage of missing values on each column on dataset 14 after
#the deletion of marathons with black
mis_perc <- colSums(is.na(data14)) / nrow(data14) * 100
mis_table <- data.frame(Missing_Percentage = mis_perc)
#sort(mis_perc)
#I have to transform CR from character to numeric. In order to do that we have
#to transform the CR to seconds (as the":" returns NA if we try to make it
#numeric just with the command "as.NA"). The CR is the best time in that
#marathon (now we will have it in seconds).
data14 <- data14 %>%
  mutate(CRseconds = sapply(strsplit(CR, ":"), function(x) {
    as.numeric(x[1]) * 3600 + as.numeric(x[2]) * 60 + as.numeric(x[3])
  }))
#Now I can calculate the variable performanceOpp (meaning Opposite). On this
#case, a higher value for performance indicates a slower time compared to the
#current course record.
data14 <- data14 %>%
  mutate(performanceOpp = data14$CRseconds * (1 + X.CR / 100))
# Calculate mean performance by age group and gender
summary_stats <- data14 %>%
  group_by(age_group, Gender) %>%
  summarise(
    mean_performance = mean(performanceOpp, na.rm = TRUE),
    median_performance = median(performanceOpp, na.rm = TRUE),
    count = n()
  )

#print(summary_stats)
#Summary statistics table sorted by mean performance
summary_table2 <- data14 %>%
  group_by(age_group, Gender) %>%
  summarise(
    mean_performance = mean(performanceOpp, na.rm = TRUE),
    median_performance = median(performanceOpp, na.rm = TRUE),
    count = n(),
    .groups = 'drop'
  ) %>%
  arrange(mean_performance)  # Sort by mean performance in ascending order

# Print summary table
# Create a nicely formatted LaTeX table for PDF
kable(summary_table2, format = "latex", booktabs = TRUE,
      caption = "Summary of Mean and Median Performance by Age Group and Gender") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down"),
                position = "center",
                full_width = FALSE) %>%
  column_spec(1:2, bold = TRUE) %>%  # Make age_group and Gender columns bold
  row_spec(0, bold = TRUE
           )  %>%
kable_styling(font_size = 7)

# Violin plot of performance by age groups and gender
ggplot(data14, aes(
  x = age_group,
  y = performanceOpp,
  fill = factor(Gender)
)) +
  geom_violin(trim = FALSE) +
  labs(title = "Violin Plot of Marathon Performance by Age Group and Gender",
       x = "Age Groups",
       y = "Performance (Lower is faster)",
       fill = "Gender") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Create a line plot to visualize the mean performance
#by age group, separated by gender.
## Mean performance plot
mean_performance_plot <- summary_stats %>%
  ggplot(aes(
    x = age_group,
    y = mean_performance,
    color = factor(Gender),
    group = Gender
  )) +
  geom_line() +
  geom_point() +
  labs(
    title = "Mean Marathon Performance by Age Group and Gender",
    x = "Age Groups",
    y = "Mean Performance (seconds - Lower is faster)",
    color = "Gender"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(mean_performance_plot)
# Histogram of performance by age groups
ggplot(data14, aes(x = performanceOpp, fill = factor(Gender))) +
  geom_histogram(binwidth = 100,
               position = "identity",
               alpha = 0.5) +
  facet_wrap( ~ age_group) +
  labs(title = "Histogram of Marathon Performance by Age Group and Gender",
       x = "Performance (Lower is faster)",
       y = "Count",
       fill = "Gender") +
  theme_minimal()


#Convert Flag and Gender variables to factor
data14$Gender <- as.factor(data14$Gender)
#levels(data14$Gender)
#how flag affect performance (with gender)#############
summary_stats <- data14 %>%
  group_by(Flag, Gender) %>%
  summarise(
    mean_performance = mean(performanceOpp, na.rm = TRUE),
    median_performance = median(performanceOpp, na.rm = TRUE),
    count = n()
  )

#print(summary_stats)
#seems like men faster than women in every Flag
##how flag affect performance (with gender)
mean_performance_plot <- data14 %>%
  group_by(Flag, Gender) %>%
  summarise(mean_performance = mean(performanceOpp, na.rm = TRUE)) %>%
  ggplot(aes(x = Flag, y = mean_performance, fill = Gender)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Mean Performance by Flag and Gender",
       x = "WBGT Flag",
       y = "Mean Performance") +
  theme_minimal()

print(mean_performance_plot)
#Lets find the relationship between Wind and Performance (and gender)
#Calculate correlation for each gender
correlations <- data14 %>%
  group_by(Gender) %>%
  summarise(correlation = cor(Wind, performanceOpp, use = "complete.obs"))

#print(correlations)
#Very weak correlation
# Creating a new dataset for means
mean_data <- data14 %>%
  group_by(Wind, Gender) %>%
  summarise(mean_performance = mean(performanceOpp, na.rm = TRUE))

ggplot(mean_data, aes(x = Wind, y = mean_performance, color =
                        as.factor(Gender))) +
  geom_line() +
  geom_point() +
  labs(title = "Mean Performance by Wind and Gender",
       x = "Wind (Speed)",
       y = "Mean Performance",
       color = "Gender") +
  theme_minimal()
#we can see that theres no clear pattern so not strong correlation (of course
#men have better performance than women)
#Lets find the relationship between Wind and Performance (and age groups)
#Calculate correlation for each gender
correlations <- data14 %>%
  group_by(age_group) %>%
  summarise(correlation = cor(Wind, performanceOpp, use = "complete.obs"))

#print(correlations)
#Very weak correlation for all age groups (a little stronger for 14 and)
# Creating a new dataset for means
mean_data <- data14 %>%
  group_by(Wind, age_group) %>%
  summarise(mean_performance = mean(performanceOpp, na.rm = TRUE))

ggplot(mean_data, aes(x = Wind, y = mean_performance, color = age_group)) +
  geom_line() +
  geom_point() +
  labs(title = "Mean Performance by Wind and Age Group",
       x = "Wind (Km/h)",
       y = "Mean Performance",
       color = "Age Group") +
  theme_minimal()
data1<-read.csv("project1.csv")
data2<-read.csv("aqi_values.csv")
data3<-read.csv("marathon_dates.csv")
data4<-read.csv("C:/Users/arontog1/Downloads/course_record (1).csv")
# Import datasets
main_data = data1
aqi_data <- read.csv("C:/Users/arontog1/Downloads/aqi_values_new.csv")
record_data = data4

# Merge main and record data sets
record_data <- record_data %>%
  mutate(Sex = ifelse(Gender == "F", 0, 1)) %>%
  mutate(Race_code = case_when(
    Race == "B" ~ 0,
    Race == "C" ~ 1,
    Race == "NY" ~ 2,
    Race == "TC" ~ 3,
    Race == "D" ~ 4
  ))

merged_main <- main_data %>%
  left_join(
    record_data[, c("Year", "Sex", "CR", "Race_code")],
    by = c(
      "Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D." = "Race_code",
      "Year" = "Year",
      "Sex..0.F..1.M." = "Sex"
    )
  ) %>%
  dplyr::rename(Race_code = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
                Sex = Sex..0.F..1.M.,
                Age = Age..yr.) %>%
  mutate(Gender = factor(ifelse(Sex == 0, "Female", "Male"))) %>%
  mutate(
    Marathon = case_when(
      Race_code == 0 ~ "Boston",
      Race_code == 1 ~ "NYC",
      Race_code == 2 ~ "Chicago",
      Race_code == 3 ~ "Twin Cities",
      Race_code == 4 ~ "Grandmas"
    )
  ) %>%
  mutate(Flag = factor(Flag, levels = c("White", "Green", "Yellow",
                                        "Red", "Black"))) %>%
  mutate(FinishTime = as.numeric(as.difftime(CR, units = "sec") * (1 + X.CR /
                                                                     100)))


#Clean up AQI by CBSA
aqi_data = aqi_data %>%
  distinct()
```

```r
aqi_mean = aqi_data %>%
  group_by(marathon, date_local, parameter, sample_duration) %>%
  summarise(daily_mean = mean(arithmetic_mean, na.rm = TRUE)) %>%
  mutate(parameter_duration = paste0(parameter, "-", sample_duration))

aqi_pivot = aqi_mean[, c("marathon", "date_local", "parameter_duration",
                         "daily_mean")] %>%
  pivot_wider(names_from = parameter_duration, values_from = daily_mean) %>%
  mutate(Year = year(date_local)) %>%
  select(
    -c(
      "PM2.5 - Local Conditions-24 HOUR",
      "Sulfur dioxide-5 MINUTE",
      "Sulfur dioxide-3-HR BLK AVG",
      "Sulfur dioxide-24-HR BLK AVG",
      "Ozone-8-HR RUN AVG BEGIN HOUR",
      "Acceptable PM2.5 AQI & Speciation Mass-1 HOUR",
      "Acceptable PM2.5 AQI & Speciation Mass-24 HOUR",
      "Acceptable PM2.5 AQI & Speciation Mass-24-HR BLK AVG",
      "PM2.5 - Local Conditions-24-HR BLK AVG"
    )
  )

merged_main = merged_main %>%
  left_join(aqi_pivot, by = c("Marathon" = "marathon", "Year" = "Year"))
names(merged_main)[names(merged_main) == "Race_code"] <- "marathon/race"
names(merged_main)[names(merged_main) == "FinishTime"] <- "PerformanceOpp"
merged_main <- select(merged_main, -date_local, -Marathon, -DP, -CR)
names(merged_main)[c(6, 7, 8, 9, 10, 11, 16, 17, 18, 19)] <- c(
  "Percent off current course record for gender",
  "Dry bulb temperature in Celsius",
  "Wet bulb temperature in Celsius",
  "Percent relative humidity",
  "Black globe temperature in Celsius",
  "Solar radiation in Watts per meter squared",
  "Nitrogen Dioxide",
  "Ozone",
  "Sulfur Dioxide",
  "PM2.5"
)
merged_main <- select(merged_main, -6)
merged_main <- merged_main %>%
  mutate(age_group = cut(
    Age,
    breaks = c(-Inf, 14, 19, 29, 39, 49, 59, 69, 79, Inf),
    labels = c(
      "14 and under",
      "15-19",
      "20-29",
      "30-39",
      "40-49",
      "50-59",
      "60-69",
      "70-79",
      "80 and over"
    ),
    right = TRUE
  ))

merged_main <- select(merged_main, -6, -7, -9)
merged_main <- merged_main[merged_main$Flag != "Black", ]
merged_main <- select(merged_main, -PM2.5)
merged_main$Age<-as.numeric(merged_main$Age)
# Calculate the percentage of missing values for each column
missingness_table <- merged_main %>%
  summarise(across(everything(), ~ sum(is.na(.)) / n() * 100)) %>%
  pivot_longer(everything(), names_to = "Variable",
               values_to = "Missing_Percentage") %>%
  arrange(desc(Missing_Percentage))
# Create a formatted table with kableExtra
kable(missingness_table, format = "latex", booktabs = TRUE,
      caption = "Percentage of Missing Values by Variable") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down"),
                position = "center",
                full_width = FALSE) %>%
  column_spec(1, bold = TRUE) %>%  # Make the "Variable" column bold
  row_spec(0, bold = TRUE) %>%
kable_styling(font_size = 7)
merged_main %>%
  dplyr::select(
    `marathon/race`,
     Sex,
     WBGT,
     Age,
     Flag,
     Wind,
    `Percent relative humidity`,
    `Solar radiation in Watts per meter squared`,
     Ozone,
    `Sulfur Dioxide`,
    `Nitrogen Dioxide`,
    PerformanceOpp
  ) %>%
  distinct() %>%
  tbl_summary(
    statistic = list(
      all_continuous() ~ c("{mean} ({sd})", "{min}, {max}"),
      all_categorical() ~ "{n} ({p}%)"
    ),
    by =`marathon/race`,  # Group by 'Marathon'
    digits = all_continuous() ~ 2,
    missing = "no",  # Exclude missing data
    type = list(
      WBGT ~ "continuous2",
      Flag ~ "categorical",
      Sex ~ "categorical",
      `Percent relative humidity` ~ "continuous2",
      Wind ~ "continuous2",
      Age ~ "continuous2",
      PerformanceOpp ~ "continuous2",
      Ozone ~ "continuous2",
       `Sulfur Dioxide` ~ "continuous2",
      `Nitrogen Dioxide` ~ "continuous2",
      `Solar radiation in Watts per meter squared`  ~ "continuous2"
    ),
    label = list(
      WBGT = "WBGT (C)",
      Sex = "Gender of the runners",
      Flag = "WBGT flag",
      `Percent relative humidity` = "Percent relative humidity",
      Wind = "Wind speed (Km/hr)",
     `Solar radiation in Watts per meter squared` = "Solar radiation (W/m^2)",
      `Sulfur Dioxide` = "SO2 pollutant ",
      `Nitrogen Dioxide` = "NO2 pollutant ",
      Ozone = "Ozone pollutant",
     Age = "Age in years",
      PerformanceOpp = "Seconds off record"
    )
  ) %>%
  modify_caption(caption =
                   "Summary of Environmental Factors by Marathon, N = {N}") %>%
  as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = "",
    format = "latex",
  ) %>%
  kableExtra::kable_styling(
    position = "center",
    latex_options = c("striped", "repeat_header"),
    stripe_color = "gray!15",
    font_size = 6
  )
# Create a data frame with the performance and environmental variables
environmental_vars <- merged_main[, c(11, 6, 7, 8, 9, 12, 13, 14)]
# Calculate the correlation matrix
correlation_matrix <- cor(environmental_vars, use = "complete.obs")
# Print the correlation coefficients with performance
performance_correlations <- correlation_matrix["PerformanceOpp", -1]
# Exclude performance itself
#print(performance_correlations)
# Identify the variable with the highest correlation
max_corr_variable <- names(performance_correlations)[which.max(abs(performance_correlations))]
max_corr_value <- max(abs(performance_correlations))
ggcorrplot(
  correlation_matrix,
  method = "circle",
  # Shape of the correlation coefficient
  type = "lower",
  lab = TRUE,
  lab_size = 2.5,
    title = "Correlation Matrix ",
  ggtheme = theme_minimal(),
  colors = c("blue", "white", "red")
) +
  theme(axis.text.x = element_text(size = 5.5),  # Adjust size of x-axis labels
        axis.text.y = element_text(size = 5.5))  # Adjust size of y-axis labels
# Scatter plot for each environmental variable vs performance, colored by
#gender and faceted by age group
env_variables <- c("Ozone",
                   "`Solar radiation in Watts per meter squared`",
                   "`Nitrogen Dioxide`", "`Sulfur Dioxide`")  #

for (var in env_variables) {
  p <- ggplot(merged_main,
              aes_string(x = var, y = "PerformanceOpp", color = "Gender")) +
    geom_point() +
    geom_smooth(method = "lm") +  # Adds a regression line
    facet_wrap( ~ age_group) +  # Facet by age group
    labs(title = paste("PerformanceOpp vs", var),
         x = var,
         y = "Performance") +
    theme_minimal() +
    theme(
      axis.text.x = element_text(size = 3),   # Smaller x-axis numbers (ticks)
      axis.text.y = element_text(size = 4),   # Smaller y-axis numbers (ticks)
      axis.title.x = element_text(size = 5), # Smaller x-axis label
      axis.title.y = element_text(size = 5), # Smaller y-axis label
      strip.text = element_text(size = 5)    # Adjust facet label size
    )

  print(p)  # Ensure ggplot is printed
}
model <- lm(
  PerformanceOpp ~ WBGT + Wind + Ozone + `Nitrogen Dioxide`
  + `Solar radiation in Watts per meter squared` ,
  data = merged_main
)

# Get the summary of the linear model
#summary(model)
model <- lm(
  PerformanceOpp ~ WBGT + Wind + Ozone + `Nitrogen Dioxide` +
    `Solar radiation in Watts per meter squared` ,
  data = merged_main
)

# Get the summary of the linear model
#summary(model)
``
```