Marathon Performance Analysis: Impact of Different Weather Conditions on Runners

William Qian

October 2024

Abstract

Introduction

Marathon running is a popular sport that attracts millions of participants worldwide. The completion time of a marathon is influenced by various factors, including weather conditions. In this report, we analyze the impact of different weather conditions on marathon performance across the lifespan in both men and women using data from five major marathons: Boston, NYC, Chicago, Twin Cities, and Grandmas.

Data Description

Two datasets are used in this analysis: Marathon Data and Course Record Data. Those two datasets were combined to form a comprehensive dataset for analysis.

Marathon Data

The Marathon Data contains information about the average completion record percentage grouped by age and sex of runners in five major marathons: Boston, NYC, Chicago, Twin Cities, and Grandmas. The dataset includes the following columns:

Parameter	Coding details
Race	0 = Boston Marathon , 1 = Chicago Marathon , 2 = New York City Marathon , 3 = Twin Cities Marathon (Minneapolis, MN) , 4 = Grandma's Marathon (Duluth, MN)
Year	Year of the marathon
Sex/Gender	0 = Identified as Female , 1 = Identified as Male
Flag	White = WBGT <10°C , Green = WBGT 10-18°C , Yellow = WBGT 18-23°C , Red = WBGT 23-28°C , Black = WBGT >28°C
% CR	Percent off current course record for gender
Td, °C	Dry bulb temperature in Celsius
Tw, °C	Wet bulb temperature in Celsius
%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

Note that the variable `WBGT` is actually a composite index that combines the effects of `Td`, `Tw`, and `Tg`, which means we can ignore the three variables when analyzing the data.

$$WBGT = 0.7 \times Tw + 0.2 \times Tg + 0.1 \times Td$$

This data contains of 11073 observations and 12 variables. The data documented the 5 major races form year 1993 to 2016, and each observation represents the average performance of a specific age in a particular race during a specific year. It should be noted that in this dataset, only one set of weather data is recorded for each race.

Course Record Data

This dataset contains the course record for each race grouped by sex documented in the Marathon Data. The dataset includes the following columns:

Parameter	Coding details
Race	B = Boston Marathon , C = Chicago Marathon , NY = New York City Marathon , TC = Twin Cities Marathon (Minneapolis, MN) , D = Grandma's Marathon (Duluth, MN)
Sex/Gender	0 = Identified as Female , 1 = Identified as Male
Year	Year of the marathon
CR	Current course record for gender

Different from the Marathon Data, the `CR` in this dataset is in hours, which means we can calculate the completion time of each runner by multiplying the CR percentage in the Marathon Data with the CR in this dataset.

Data Preprocessing

Data Manipulation

For clarity, we firstly rename the `Race` column in the Marathon Data as well as the Course Record Data to the full name of the marathon. And since we are using two datasets, we need to merge them based on the `Race`, `Year`, and `Sex` columns. After merging, we now have a united dataset that contains 15 variables and 11073 observations in total.

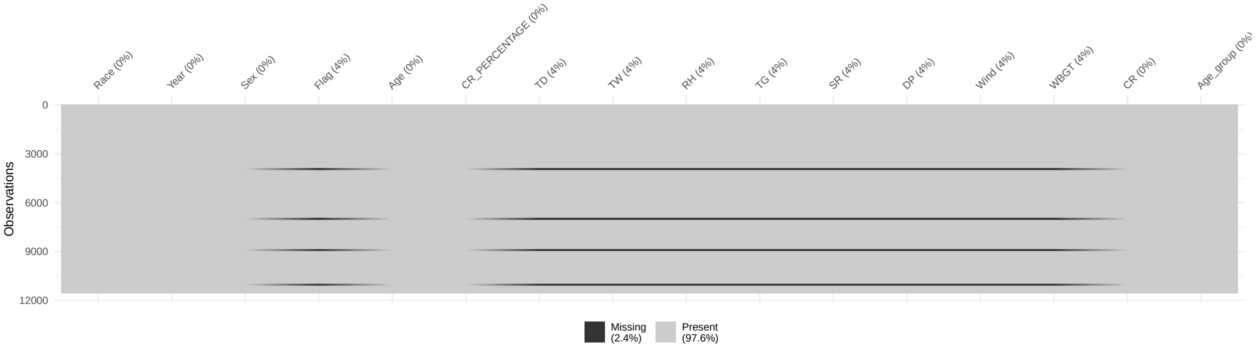
And then, condiering of the nature of the variables and the analysis we are going to perform, we convert the `Year`, `Race`, `Sex`, and `Flag` columns to factors. We also created a new column `Age_group` by grouping the `Age` column into 10-year intervals. This step is essential trying to factorize the `Age` column.

Most importantly, we would like to convert the `CR` in the Course Record Data to seconds. The process is simple, we first convert the `CR` to a period object and then convert it to seconds. After that, we can calculate the completion time of each runner by multiplying the CR percentage in the Marathon Data with the CR in this dataset. The calculation formula is as follows:

$$CR_{adjusted} = (1 + CR_{PERCENTAGE} \times 0.01) \times CR$$

Data Quality Check

Before we start the analysis, we need to check the data quality. We first check for missing values and patterns in the dataset. The missing values are visualized using the `vis_miss` function from the `naniar` package. The plot shows that there are missing values in the weather related columns, and this missingness seems to have a very clear pattern. Based on this pattern, we can assume boldly that the missing values are not missing at random, but are related to specific races.

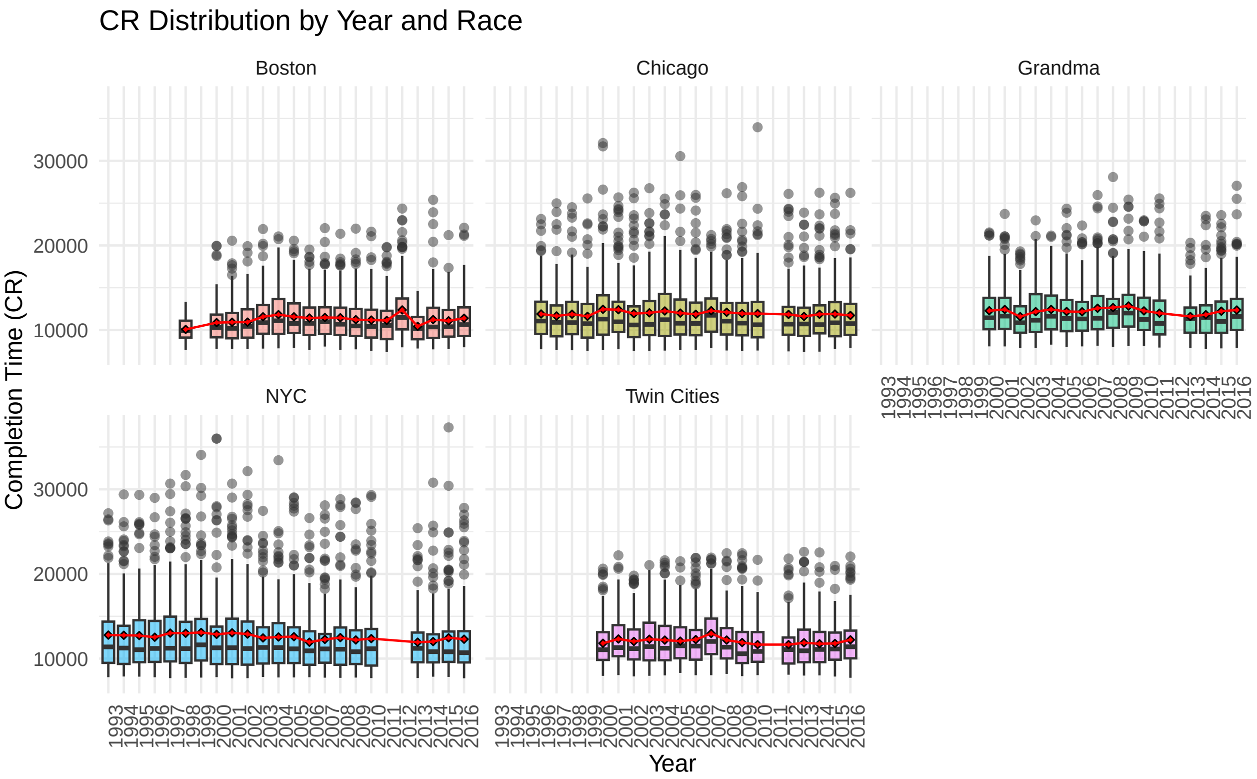


In order to veryify our assumption, we calculate the missing percentage of weather data in each marathon by year. The table below shows that the missing percentage of most of the races are 0, however, for races held in 2011 (Chicago, NYC, and Twin Cities) as well as the one held in 2012 (Grandmas), the missing percentage is 100%. This confirms our assumption that the missing values are related to specific races. It would be a wise choice to remove all of those races from our dataset. To be noted that the NA value in the table means there are no races held in that year.

Table 3: Missing Percentage of Weather Data in Each Race by Year

Year	Boston	Chicago	Grandma	NYC	Twin Cities
1993	0	0	NA	0	0
1994	0	0	NA	0	0
1995	0	0	NA	0	0
1996	0	0	NA	0	0
1997	0	0	NA	0	0
1998	0	0	NA	0	0
1999	0	0	NA	0	0
2000	0	0	0	0	0
2001	0	0	0	0	0
2002	0	0	0	0	0
2003	0	0	0	0	0
2004	0	0	0	0	0
2005	0	0	0	0	0
2006	0	0	0	0	0
2007	0	0	0	0	0
2008	0	0	0	0	0
2009	0	0	0	0	0
2010	0	0	0	0	0
2011	0	1	0	1	1
2012	0	0	1	0	0
2013	0	0	0	0	0
2014	0	0	0	0	0
2015	0	0	0	0	0
2016	0	0	0	0	0

And we also want to check the performance distribution of each race by year, since although the weather condition may vary from year to year, the performance of runners may not change significantly within the same track. As we can see in the plot below, despite some disturbances, the performance of runners tend to stay stable over the years. The difference in performance might be due to the different weather conditions in different years. Overall, the stability of the performance indicates that the dataset is reliable for analysis.



Another factor might affect the result of the analysis is `age`. In a balanced dataset, we would expect the number of participants in each age group to be roughly the same. To check this, we count the number of participants in each age group for each races. The table shows a pattern that for each race, participants between age 20 to 79 are the most common and also balanced, while the number of participants in the 10-19, 80-89 and 90-99 age groups are relatively small. This suggests that the inference results obtained from the 10-19, 80-89 and 90-99 age groups may not be as reliable as those obtained from the 20-79 age groups. Moreover, the number of participants in the 90-99 age group is extremely small, so we decide to merge it with the 80-89 age group to become the 80-99 age group.

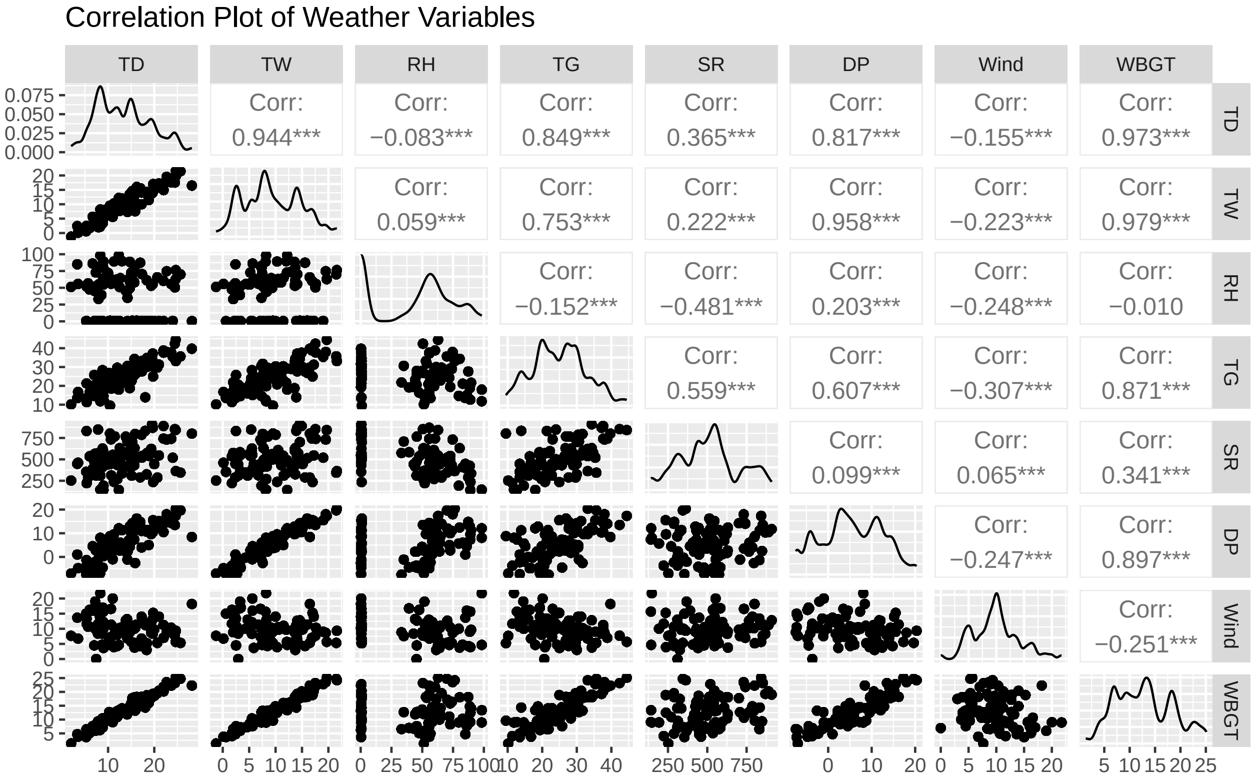
Similarly, we would hope that participants' gender is balanced in the dataset. From the table below, we can tell the Female:Male portion is roughly equal to 1 in all of the races, which is a good feature for our analysis.

Table 4: **Number of Participants by Age Group, Sex and Race**

Age Group	**Boston**, N = 2,088	**Chicago**, N = 2,427	**Grandma**, N = 1,884	**NYC**, N = 2,799	**Twin Cities**, N = 1,875
Age_group					
0-9	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
10-19	67 (3.2%)	171 (7.0%)	167 (8.9%)	88 (3.1%)	140 (7.5%)
20-29	360 (17%)	400 (16%)	320 (17%)	440 (16%)	319 (17%)
30-39	360 (17%)	400 (16%)	320 (17%)	440 (16%)	320 (17%)
40-49	360 (17%)	400 (16%)	320 (17%)	440 (16%)	320 (17%)
50-59	359 (17%)	400 (16%)	320 (17%)	440 (16%)	319 (17%)
60-69	337 (16%)	391 (16%)	286 (15%)	438 (16%)	294 (16%)
70-79	215 (10%)	237 (9.8%)	134 (7.1%)	378 (14%)	146 (7.8%)
80-89	30 (1.4%)	28 (1.2%)	17 (0.9%)	130 (4.6%)	17 (0.9%)
90-99	0 (0%)	0 (0%)	0 (0%)	5 (0.2%)	0 (0%)
Sex					
Female	984 (47%)	1,150 (47%)	880 (47%)	1,337 (48%)	867 (46%)

Age Group	**Boston**, N = 2,088	**Chicago**, N = 2,427	**Grandma**, N = 1,884	**NYC**, N = 2,799	**Twin Cities**, N = 1,875
Male	1,104 (53%)	1,277 (53%)	1,004 (53%)	1,462 (52%)	1,008 (54%)

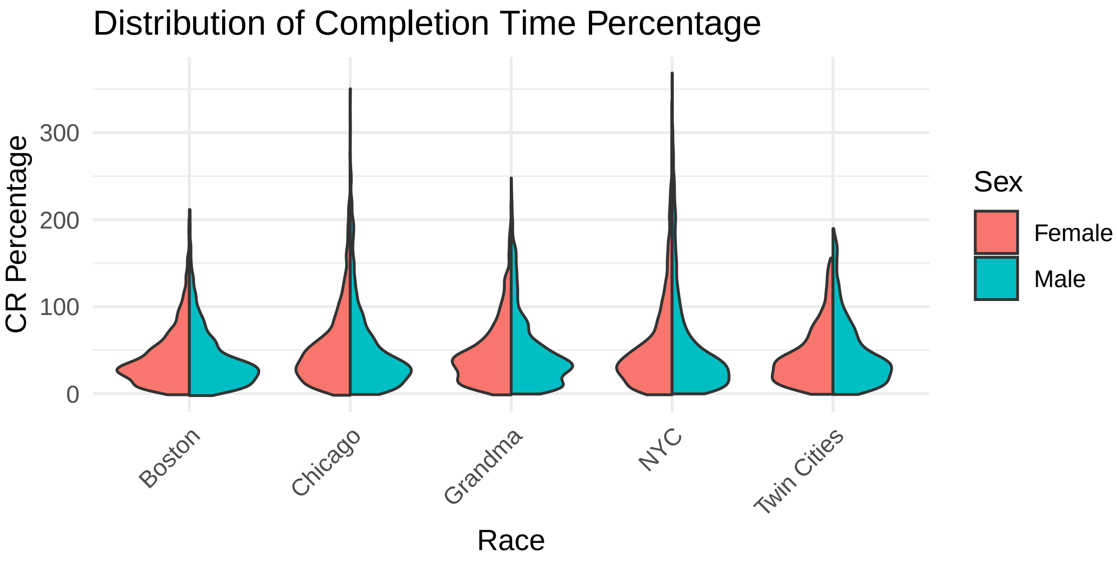
We also want to check the correlation between the weather variables. The correlation plot below shows that `DP` is highly correlated with `WBGT`, which means we can ignore the `DP` variable when analyzing the data. And since the `WBGT` is a composite index that combines the effects of `Td`, `Tw`, and `Tg`, all of those variables are highly correlated with `WBGT`, meaning that they can be ignored as well.



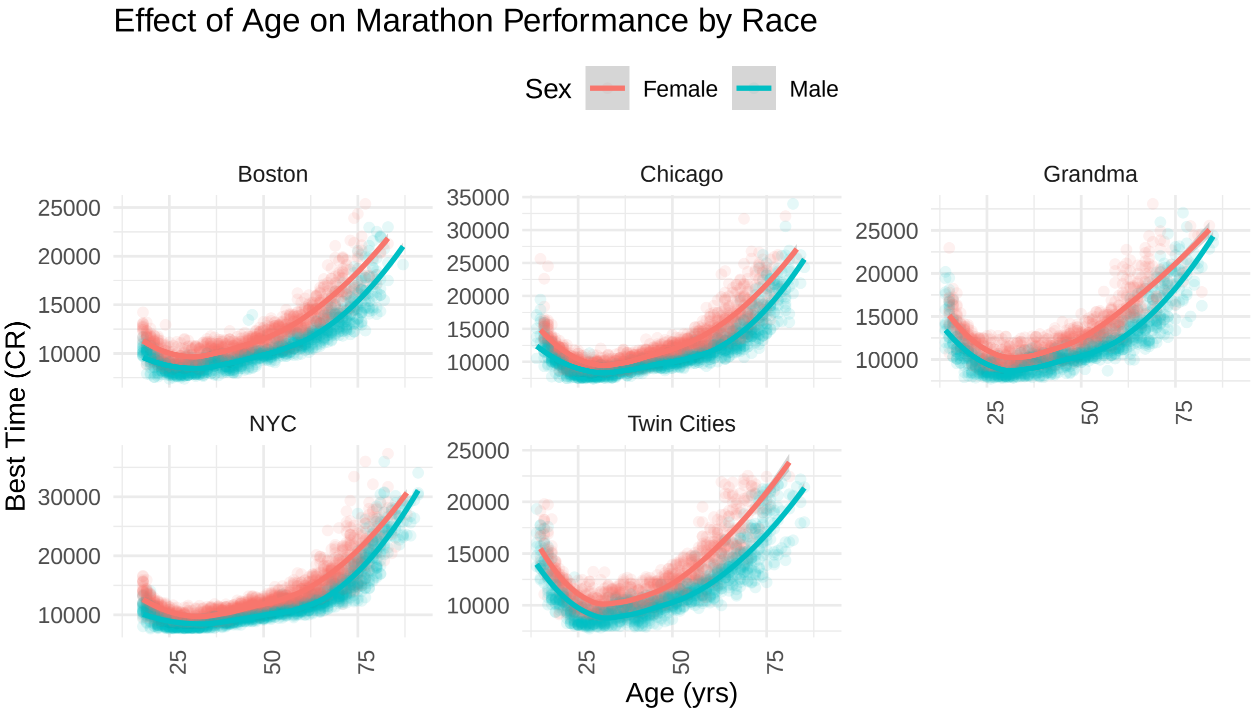
Data Analysis

Inner Factors Analysis

A very instinctive thought is that the completion time of a marathon is influenced by gender. To verify this, we first plot the distribution of completion time percentage among Female and Male, and find that in each races, males tend to have a lower completion time. And we can also observe a distribution pattern that the completion time for both male and female runners is heavily skewed, which means that most of the runners have a completion time close to the course record. This suggests that our runners are generally well-trained and have a good performance.



Next, we want to explore the effect of age on marathon performance. From the plot below, we can not only observe the fact that female runners are tend to run slower than male runners, but we can also observe that in each cases, no matter male or female, the completion time decreases first and then increases with age. And people around their 30s tend to have the best performance. This suggests that the completion time of a marathon is influenced by age, and the effect of age on completion time is not linear.



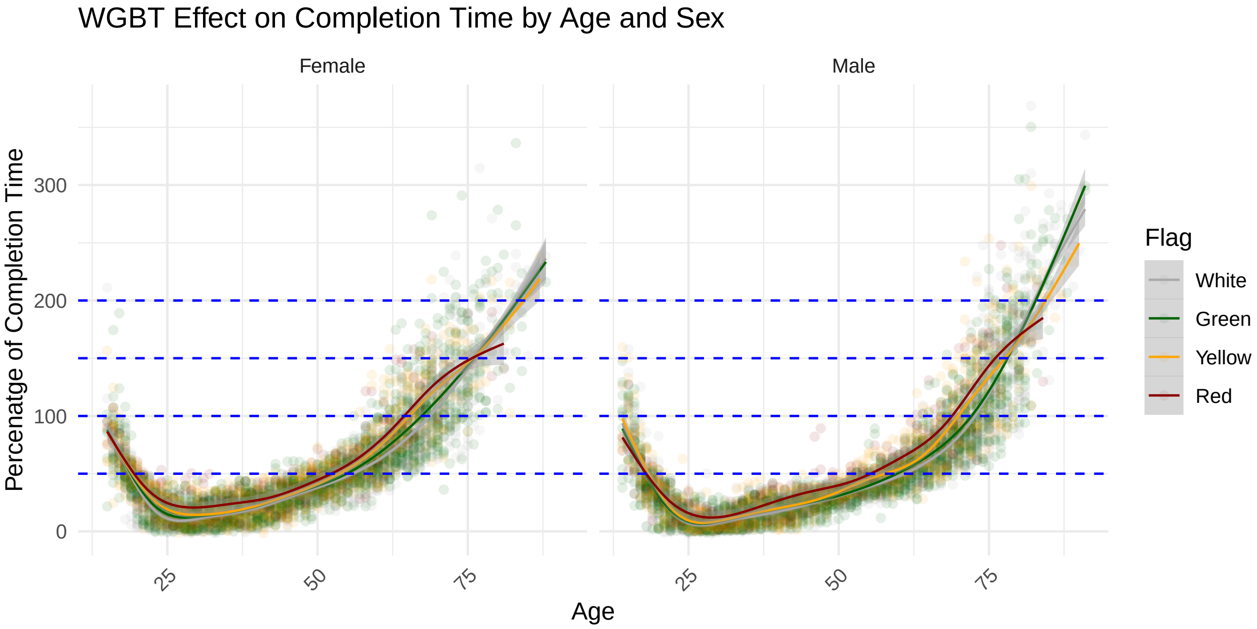
Weather (Outer) Factors Analysis

Both of `Age` and `Sex` can be considered as inner factors that affect the completion time of a marathon. We also care about the outer factors such as weather condition. Multiple weather variables are recorded in the dataset, including `RH`, `SR`, `DP`, `Wind`, and `WBGT`. In the following section, we will analyze the effect of these weather variables on marathon performance, and also explore the interaction between weather variables and inner factors.

We first want to focus on the effect of WBGT on marathon performance. WBGT is a composite index that combines the effects of Td, Tw, and Tg. Conceptually, the WBGT is an index used to estimate the effect of temperature, humidity, wind speed, and solar radiation on humans, typically to assess heat stress during physical activities in outdoor environments. That suggests the higher the WBGT, the more difficult for runners to finish the marathon.

In order to prove our assumption, we plot the effect of WBGT on completion time by sex. However, instead of using WBGT directly, we use the Flag column to represent the WBGT level. The Flag column is a categorical variable that represents the WBGT level, with White representing the lowest WBGT level and Black representing the highest WBGT level. In the following plot, each of the plot represents an observation, and the color of the plot represents the Flag level. To better read the result, we also add linear regression line to each plot.

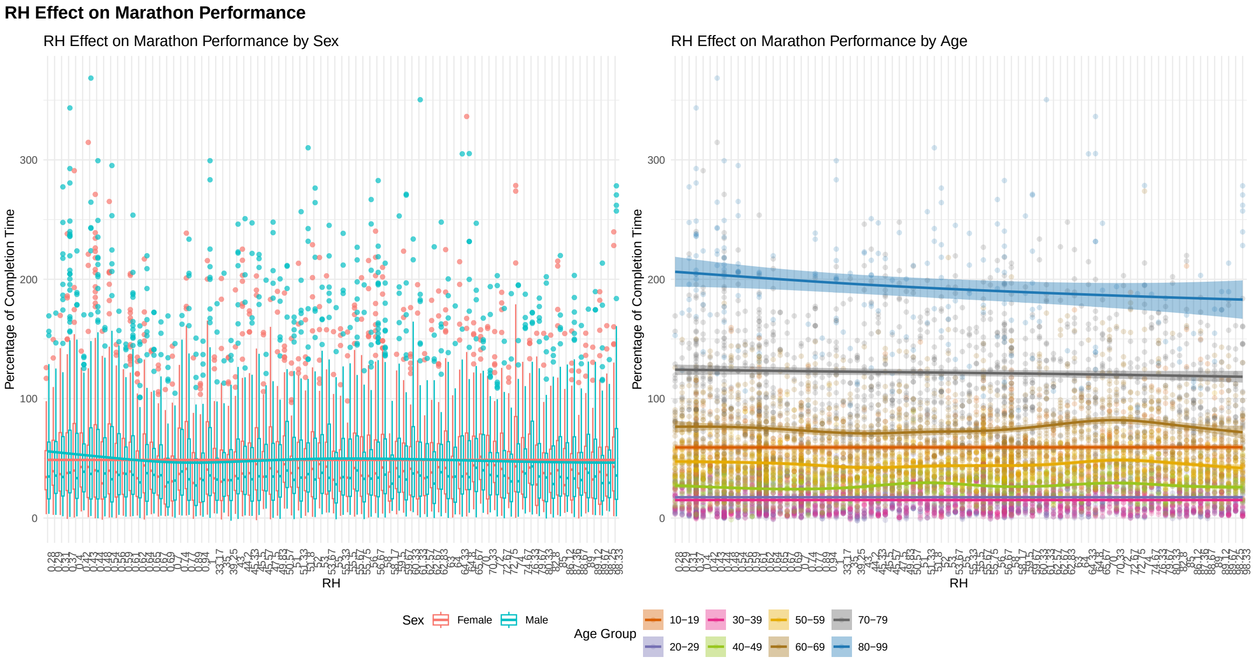
By interpreting the plot, we can tell that in general cases, the worse the WBGT level, the higher the completion time. In detail, we found that the WBGT level do have different effect for different age groups as well as gender. To be note that, the wider the space between two regression lines, the more significant the effect of WBGT level on completion time. In that case, we can tell the WBGT level has a more significant effect on male than female. And looking at the difference among age groups, we found that the WBGT has more effect on elder runners than younger runners. Additionaly, we see intersection between the regression lines around age 75, we think that might be due to the small number of participants in the 70-79 age group, causing the result to be less reliable, the larger standard error of the regression line showing in the plot also suggests this assumption.

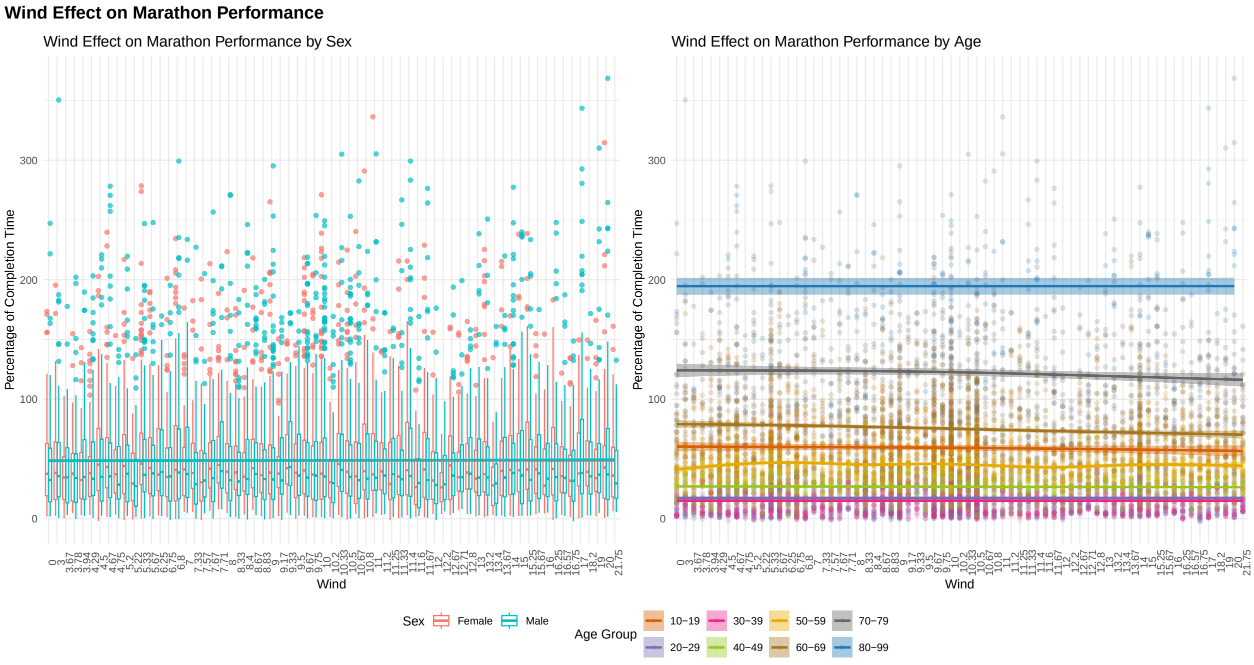


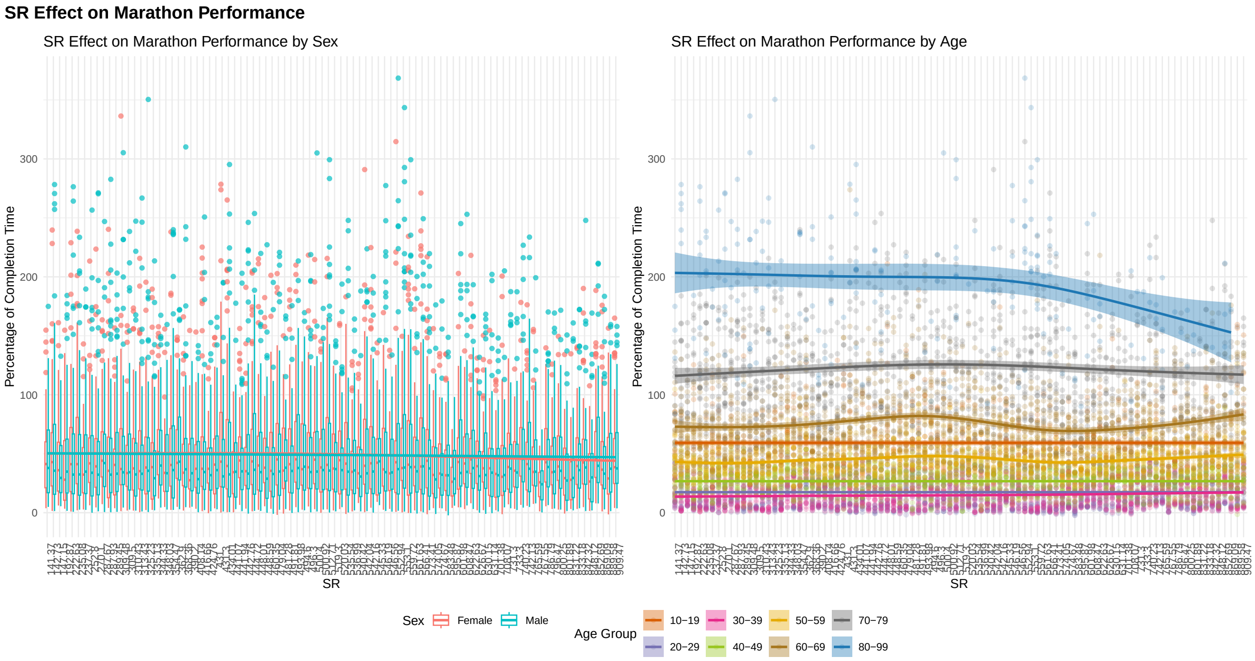
Next, let's check the rest of the weather variables. RH is the relative humidity, which is the ratio of the amount of water vapor present in the air to the maximum amount that the air can hold at that temperature. SR is the solar radiation, which is the power received per unit area from the sun in the form of electromagnetic radiation in the wavelength range of the solar spectrum. Wind is the wind speed, which is the rate at which air flows past a point on the earth's surface in a unit of time.

Since all of those variables are continuous, we set two plots for each of them. The first plot is the boxplot used to show the effect on gender. The weather variable is set as the x-axis, and the completion time percentage is set as the y-axis. The color of the box represents different gender. The second plot is the scatter plot used to show the effect on age. The setting of x-axis and y-axis is same as the first plot. The color of the point represents different age groups. In both of the plots, we also add a linear regression line to show the trend of the data, the slope of the regression line represents the effect of the weather variable on completion time.

For the RH variable, we observed a roughly flat trend for both male and female runners, which means the RH level has little effect on completion time. And the gap between the regression lines for different gender groups is relatively small, which means the effect of RH level on completion time is similar for different gender. For different age groups, we found the trend stays flat for most of the age groups, except for the 80-99 age group, which has a slightly decreasing trend. This suggests the higher the age, the less effect the RH level has on completion time, which is not matching with common sense. This might be due to the small number of participants in the 80-99 age group, causing the result to be deviated. For the SR and Wind variable, we observed a similar pattern in RH.







Model Building

The visualization results is a good way to show the general effect or the trend of the data, however, it is not a good way to quantify the effect of the weather variables on performance. In order to quantify the effect, we use linear regression models to reflect the effect. Several models were build.

Firstly, we build a model that includes the Sex and Age variables. The result shows that the Sex variable has a negative effect of -4.7812 on completion percentage, while the Age variable has a positive effect of 1.7545 on completion percentage. Both of the P-values are significant. The estimates means that males are expected to have a 4.7812 percent faster completion than female, and the completion percentage will increase by 1.7545 percent for each year increase in the Age variable.

Table 5: Sex and Age Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-30.215186	0.8782862	-34.402438	0
SexMale	-4.781230	0.6080130	-7.863696	0
Age	1.754557	0.0168972	103.836895	0

$$\begin{aligned} Model_{sex\_age} : CR_{PERCENTAGE} &= \beta_0 + \beta_1 \times Sex + \beta_2 \times Age \\ &= -30.2151 - 4.7812 \times Sex_{male} + 1.7545 \times Age \end{aligned}$$

The RH model, which includes the RH, Sex, and Age variables as well as the interaction terms between RH and Sex, and RH and Age. The result shows that the RH variable has a positive effect of 0.1637 on completion percentage. Also, the Sex variable has a negative effect of -5.3774 on completion percentage, while the Age variable has a positive effect of 1.9105 on completion percentage. The P-values of all of the estimates are significant. The interaction term of RH and Sex has a positive effect of 0.0128 on completion percentage, while the interaction term of RH and Age has a negative effect of -0.0036 on completion percentage. The P-value of RH-Sex interaction term is 0.4983, which is not significant, suggesting that the RH effect does not differ between male and female runners. The estimate of RH-Age interaction term is significant, suggesting that the effect of RH on completion percentage is different for different age groups, more specifically, with same level of RH, runners will have an extra 0.0036 percent decrease for each year increase in the Age.

Table 6: RH Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-37.1520990	1.4553234	-25.5284152	0.0000000
RH	0.1637750	0.0273895	5.9794740	0.0000000
SexMale	-5.3774050	1.0078170	-5.3356960	0.0000001
Age	1.9105133	0.0280386	68.1387649	0.0000000
RH:SexMale	0.0128486	0.0189750	0.6771334	0.4983355
RH:Age	-0.0036852	0.0005293	-6.9624974	0.0000000

$$\begin{aligned} Model_{RH} : CR_{PERCENTAGE} &= \beta_0 + \beta_1 \times RH + \beta_2 \times Sex + \beta_3 \times Age + \beta_4 \times RH \times Sex + \beta_5 \times RH \times Age \\ &= \beta_0 + \beta_2 \times Sex + \beta_3 \times Age + (\beta_1 + \beta_4 \times Sex + \beta_5 \times Age) \times RH \\ &= -37.1520 - 5.3774 \times Sex + 1.9105 \times Age + (0.1637 + 0.0128 \times Sex - 0.0036 \times Age) \times RH \end{aligned}$$

The SR model, which includes the SR, Sex, and Age variables as well as the interaction terms between SR and Sex, and SR and Age. The result shows that the SR variable has a positive effect of 0.0271 on completion percentage. Also, the Sex variable has a negative effect of -6.1501 on completion percentage, while the Age variable has a positive effect of 2.0834 on completion percentage. The P-values of all of the estimates are significant. The interaction term of SR and Sex has a positive effect of 0.0027 on completion percentage, while the interaction term of SR and Age has a negative effect of -0.0006 on completion percentage. The P-value of SR-Sex interaction term is 0.3946, which is not significant, suggesting that the SR effect does not differ between male and female runners. The estimate of SR-Age interaction term is significant, suggesting that the effect of SR on completion percentage is different for different age groups, more specifically, with same level of SR, runners will have an extra 0.0006 percent decrease for each year increase in the Age.

Table 7: SR Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-44.0394398	2.5607094	-17.1981401	0.0000000
SR	0.0271455	0.0046829	5.7967130	0.0000000
SexMale	-6.1501991	1.7682169	-3.4781928	0.0005067
Age	2.0834985	0.0489854	42.5330131	0.0000000
SR:SexMale	0.0027535	0.0032347	0.8512602	0.3946433
SR:Age	-0.0006503	0.0000907	-7.1669548	0.0000000

$$\begin{aligned} Model_{SR} : CR_{PERCENTAGE} &= \beta_0 + \beta_1 \times SR + \beta_2 \times Sex + \beta_3 \times Age + \beta_4 \times SR \times Sex + \beta_5 \times SR \times Age \\ &= \beta_0 + \beta_2 \times Sex + \beta_3 \times Age + (\beta_1 + \beta_4 \times Sex + \beta_5 \times Age) \times SR \\ &= -44.0394 - 6.1501 \times Sex + 2.0834 \times Age + (0.0271 + 0.0027 \times Sex - 0.0006 \times Age) \times SR \end{aligned}$$

The Wind model, which includes the Wind, Sex, and Age variables as well as the interaction terms between Wind and Sex, and Wind and Age. The result shows that the Wind variable has a negative effect of -1.0906 on completion percentage. Also, the Sex variable has a negative effect of -6.2381 on completion percentage, while the Age variable has a positive effect of 1.5773 on completion percentage. The P-values of all of the estimates are significant. The interaction term of Wind and Sex has a positive effect of 0.1449 on completion percentage, while the interaction term of `Wind` and `Age` has a positive effect of 0.01785 on completion percentage. The P-value of `Wind-Sex` interaction term is 0.3314, which is not significant, suggesting that the `Wind` effect does not differ between male and female runners. The estimate of `Wind-Age` interaction term is significant, suggesting that the effect of `Wind` on completion percentage is different for different age groups, more specifically, with same level of `Wind`, runners will have an extra 0.0178 percent increase for each year increase in the `Age`.

Table 8: Wind Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.4181455	2.3291616	-8.3369678	0.0000000
Wind	-1.0906781	0.2176067	-5.0121527	0.0000005
SexMale	-6.2381655	1.6032403	-3.8909734	0.0001004
Age	1.5773520	0.0447314	35.2627148	0.0000000
Wind:SexMale	0.1449589	0.1492629	0.9711649	0.3314874
Wind:Age	0.0178579	0.0041498	4.3033489	0.0000170

$$\begin{aligned} Model_{Wind} : CR_{PERCENTAGE} &= \beta_0 + \beta_1 \times Wind + \beta_2 \times Sex + \beta_3 \times Age + \beta_4 \times Wind \times Sex + \beta_5 \times Wind \times Age \\ &= \beta_0 + \beta_2 \times Sex + \beta_3 \times Age + (\beta_1 + \beta_4 \times Sex + \beta_5 \times Age) \times Wind \\ &= -19.4181 - 6.2381 \times Sex + 1.5773 \times Age + (-1.0906 + 0.14495 \times Sex + 0.0178 \times Age) \times \end{aligned}$$

The `WBGT` model, which includes the `WBGT`, `Sex`, and `Age` variables as well as the interaction terms between `WBGT` and `Sex`, and `WBGT` and `Age`. The result shows that the `WBGT` variable has a positive effect of 1.2692 on completion percentage. Also, the `Sex` variable has a negative effect of -4.4349 on completion percentage, while the `Age` variable has a positive effect of 1.9852 on completion percentage. The P-values of all of the estimates are significant. The interaction term of `WBGT` and `Sex` has a negative effect of -0.0286 on completion percentage, while the interaction term of `WBGT` and `Age` has a negative effect of -0.0174 on completion percentage. The P-value of `WBGT-Sex` interaction term is 0.7915, which is not significant, suggesting that the `WBGT` effect does not differ between male and female runners. The estimate of `WBGT-Age` interaction term is significant, suggesting that the effect of `WBGT` on completion percentage is different for different age groups, more specifically, with same level of `WBGT`, runners will have an extra 0.0174 percent decrease for each year increase in the `Age`.

Table 9: WBGT Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-46.8933940	2.2028065	-21.2880227	0.0000000
WBGT	1.2692664	0.1555470	8.1600200	0.0000000
SexMale	-4.4349396	1.5240973	-2.9098796	0.0036229
Age	1.9852210	0.0422178	47.0233706	0.0000000
WBGT:SexMale	-0.0286229	0.1083180	-0.2642484	0.7915934
WBGT:Age	-0.0174927	0.0030175	-5.7971717	0.0000000

$$\begin{aligned} Model_{WBGT} : CR_{PERCENTAGE} &= \beta_0 + \beta_1 \times WBGT + \beta_2 \times Sex + \beta_3 \times Age + \beta_4 \times WBGT \times Sex + \beta_5 \times WBGT \times Age \\ &= \beta_0 + \beta_2 \times Sex + \beta_3 \times Age + (\beta_1 + \beta_4 \times Sex + \beta_5 \times Age) \times WBGT \\ &= -46.8933 - 4.4349 \times Sex + 1.9852 \times Age + (1.2692 - 0.0286 \times Sex - 0.0174 \times Age) \times W \end{aligned}$$

Table 10: Model with WBGT Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-31.7106499	1.8798600	-16.8686229	0.0000000
RH	-0.0247545	0.0111761	-2.2149539	0.0267835
SR	-0.0089618	0.0019884	-4.5070711	0.0000066
Wind	-0.0059337	0.0792797	-0.0748458	0.9403388
WBGT	0.5475309	0.0604373	9.0594912	0.0000000
SexMale	-4.7856304	0.6057137	-7.9008130	0.0000000
Age	1.7575322	0.0168700	104.1808198	0.0000000

Table 11: Model with Flag Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28.4170798	1.8871588	-15.0581287	0.0000000
RH	-0.0255457	0.0116163	-2.1991279	0.0278894
SR	-0.0085841	0.0019736	-4.3494429	0.0000138
Wind	0.0364732	0.0831440	0.4386747	0.6609058
FlagGreen	3.1879712	0.7622314	4.1824190	0.0000291
FlagYellow	6.6874747	0.9459393	7.0696655	0.0000000
FlagRed	12.4054641	1.5509085	7.9988367	0.0000000
SexMale	-4.7811319	0.6055593	-7.8953981	0.0000000
Age	1.7565324	0.0168651	104.1519500	0.0000000

Discussion

References

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
library(mice, warn.conflicts = FALSE)
library(naniar)
library(ggplot2)
library(dplyr)
library(readr)
library(tidyr)
library(readxl)
library(ggpubr)
library(gtsummary)
library(GGally)
library(ggcorrplot)
library(knitr)
library(kableExtra)
library(lubridate)
library(patchwork)
library(introdataviz)
```

```r
# Load data
marathon_data <- read.csv("../Data/project1.csv")
course_record <- read.csv("../Data/course_record.csv")

# rename the column names that are too long to follow.
colnames(marathon_data)[1] <- "Race"
colnames(marathon_data)[3] <- "Sex"
colnames(marathon_data)[5] <- "Age"
colnames(marathon_data)[6] <- "CR_PERCENTAGE"
colnames(marathon_data)[7] <- "TD"
colnames(marathon_data)[8] <- "TW"
colnames(marathon_data)[9] <- "RH"
colnames(marathon_data)[10] <- "TG"
colnames(marathon_data)[11] <- "SR"

# data type conversion
marathon_data$Year <- as.factor(marathon_data$Year)
marathon_data$Race <- as.factor(marathon_data$Race)
marathon_data$Sex <- as.factor(marathon_data$Sex)
marathon_data$Flag <- as.factor(marathon_data$Flag)

marathon_data$Flag[marathon_data$Flag == ""] <- NA

# replace marathon name with code name in course_record
course_record$Race[course_record$Race == "B"] <- 0
course_record$Race[course_record$Race == "C"] <- 1
course_record$Race[course_record$Race == "NY"] <- 2
course_record$Race[course_record$Race == "TC"] <- 3
course_record$Race[course_record$Race == "D"] <- 4
course_record$Race <- as.factor(course_record$Race)

# replace gender in course_record
course_record$Gender[course_record$Gender == "M"] <- 1
course_record$Gender[course_record$Gender == "F"] <- 0
course_record$Gender <- as.factor(course_record$Gender)
colnames(course_record)[4] <- "Sex"

# Transform records in course_record into seconds
course_record$CR <- period_to_seconds(hms(course_record$CR))

# Join course_record and marathon_data
marathon_data <- merge(marathon_data, course_record, by = c("Race", "Year", "Sex"))

# calculate the record of each runner
marathon_data$CR <- (1 + marathon_data$CR_PERCENTAGE * 0.01) * marathon_data$CR

marathon_data <- marathon_data %>%
  mutate(Race = case_when(
    Race == 0 ~ "Boston",
    Race == 1 ~ "Chicago",
    Race == 2 ~ "NYC",
    Race == 3 ~ "Twin Cities",
    Race == 4 ~ "Grandma"
  ),
  Sex = case_when(
    Sex == 1 ~ "Male",
    Sex == 0 ~ "Female"
  )) %>%
  mutate(Age_group = cut(Age, breaks = seq(0, 100, by = 10), right = FALSE,
                         labels = c("0-9", "10-19", "20-29", "30-39", "40-49",
                                    "50-59", "60-69", "70-79", "80-89", "90-99")))

marathon_data$Flag <- factor(marathon_data$Flag, levels = c("White", "Green", "Yellow", "Red", "Black")

# Check for missing values and patterns
miss_plot <- vis_miss(marathon_data)
ggsave("../Plots/missing_values_plot.png", plot = miss_plot, width = 15, dpi = 300)
miss_plot

# Check the missing percentage of weather data in each marathon by year
marathon_data %>%
  group_by(Race, Year) %>%
  summarise(missing_percentage = sum(is.na(Flag)) / n()) %>%
  pivot_wider(names_from="Race", values_from = missing_percentage) %>%
  arrange(Year) %>%
  replace_na(list(Boston = 0, Chicago = 0, NYC = 0, `Twin Cities` = 0, Grandmas = 0)) %>%
  kable(caption = "Missing Percentage of Weather Data in Each Race by Year")

# remove missing data
marathon_data <- marathon_data %>% filter(!is.na(Flag))

completion_time_race <- ggplot(marathon_data, aes(x = Year, y = CR)) +
  geom_boxplot(aes(fill = Race), alpha = 0.5) +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 1, fill = "red") +
  stat_summary(fun = "mean", geom = "line", aes(group = 1), color = "red") +
  facet_wrap(~ Race) +
  labs(title = "CR Distribution by Year and Race",
       x = "Year",
       y = "Completion Time (CR)") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1),
    legend.position = "none"
  )
ggsave("../Plots/completion_time_race.png", plot = completion_time_race, width = 10, height = 8, dpi = 3
completion_time_race

tbl_summary(
  marathon_data %>% select(Race, Age_group, Sex),
  by = Race,
    statistic = list(
    Age_group ~ "{n} ({p}%)",
    Sex ~ "{n} ({p}%)"
  )
) %>%
  modify_header(label = "**Age Group**") %>%
  modify_caption("**Number of Participants by Age Group, Sex and Race**")
participants_age_plot <- ggplot(marathon_data, aes(x = Age_group, fill = Age_group)) +
  geom_bar() +
  facet_wrap(~ Race) +
  scale_fill_viridis_d() +
  labs(title = "Number of Participants by Age Group for Each Race",
       x = "Age Group",
       y = "Number of Participants",
       fill = "Age Group") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggsave("../Plots/participants_age_plot.png", plot = participants_age_plot, width = 10, height = 8, dpi =
participants_age_plot

marathon_data <- marathon_data %>%
  mutate(Age_group = if_else(Age_group == "90-99", "80-99", Age_group)) %>%
  mutate(Age_group = if_else(Age_group == "80-89", "80-99", Age_group))
sex_distribution_race <- ggplot(marathon_data, aes(x = Sex, fill = Sex)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  facet_wrap(~ Race, scales = "free_y") +
  labs(title = "Sex Distribution by Race",
       x = "Sex",
       y = "Count",
       fill = "Sex") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(face = "italic"),
    legend.position = "none"
  )
ggsave("../Plots/sex_distribution_race.png", plot = sex_distribution_race, width = 10, height = 8, dpi =
sex_distribution_race

cor_plot <- ggpairs(marathon_data %>% select(TD, TW, RH, TG, SR, DP, Wind, WBGT)) + ggtitle("Correlation
ggsave("../Plots/cor_plot.png", plot = cor_plot, width = 10, height = 8, dpi = 300)
cor_plot

cr_distribution_plot <- ggplot(marathon_data, aes(x=Race, y = CR_PERCENTAGE, fill=Sex)) +
  geom_split_violin() +
  labs(title = "Distribution of Completion Time Percentage",
       y = "CR Percentage") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
ggsave("../Plots/cr_distribution_plot.png", plot = cr_distribution_plot, width = 8, height = 5, dpi = 3
cr_distribution_plot

age_effect_plot <- ggplot(marathon_data, aes(x = Age, y = CR, color = Sex)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "loess", se = TRUE) +
  facet_wrap(~ Race, scales = "free_y") +
  labs(title = "Effect of Age on Marathon Performance by Race",
       x = "Age (yrs)",
       y = "Best Time (CR)",
       color = "Sex") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1),
    legend.position = "top"
  )
ggsave("../Plots/age_effect_plot.png", plot = age_effect_plot, width = 10, height = 8, dpi = 300)
age_effect_plot

# WGBT effects
flag_colors <- c("Green" = "darkgreen",
                 "Yellow" = "orange",
                 "Red" = "darkred",
                 "White" = "darkgrey")


wgbt_effect <- marathon_data %>%
  ggplot(aes(x = Age, y = CR_PERCENTAGE, color = Flag)) +
  facet_wrap(~ Sex, scales = "fixed") +
  geom_point(alpha = 0.1) +
  geom_smooth(aes(group = Flag, color = Flag), se = T, size=0.5) +
  geom_hline(yintercept = 50, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 100, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 150, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 200, linetype = "dashed", color = "blue") +
  scale_color_manual(values = flag_colors) +
  labs(title = "WGBT Effect on Completion Time by Age and Sex",
       x = "Age",
       y = "Percenatge of Completion Time",
       color = "Flag") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
ggsave("../Plots/wgbt_effect.png", plot = wgbt_effect, width = 10, height = 8, dpi = 300)
wgbt_effect

age_group_colors <- c("0-9" = "#1b9e77",
                      "10-19" = "#d95f02",
                      "20-29" = "#7570b3",
                      "30-39" = "#e7298a",
                      "40-49" = "#98c61e",
                      "50-59" = "#e6ab02",
                      "60-69" = "#a6761d",
                      "70-79" = "#666666",
                      "80-99" = "#1f78b4")

# RH Effects
RH_sex <- marathon_data %>%
      ggplot(aes(x = as.factor(round(RH, 2)), y = CR_PERCENTAGE, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
```

```r
      geom_smooth(aes(group = Sex), se = FALSE) +
      labs(title = "RH Effect on Marathon Performance by Sex",
           x = "RH",
           y = "Percentage of Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

RH_age <- marathon_data %>%
      ggplot(aes(x = as.factor(round(RH, 2)), y = CR_PERCENTAGE, fill = Age_group, color = Age_group))
      geom_point(alpha = 0.2) +
      geom_smooth(aes(group = Age_group, color = Age_group), se = T) +
      scale_fill_manual(values = age_group_colors) +
      scale_color_manual(values = age_group_colors) +
      labs(title = "RH Effect on Marathon Performance by Age",
           x = "RH",
           y = "Percentage of Completion Time",
           fill = "Age Group",
           color = "Age Group") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

# Wind Effects
Wind_sex <- marathon_data %>%
      ggplot(aes(x = as.factor(round(Wind, 2)), y = CR_PERCENTAGE, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = Sex), se = FALSE) +
      labs(title = "Wind Effect on Marathon Performance by Sex",
           x = "Wind",
           y = "Percentage of Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

Wind_age <- marathon_data %>%
      ggplot(aes(x = as.factor(round(Wind, 2)), y = CR_PERCENTAGE, fill = Age_group, color = Age_group)
      geom_point(alpha = 0.2) +
      geom_smooth(aes(group = Age_group, color = Age_group), se = T) +
      scale_fill_manual(values = age_group_colors) +
      scale_color_manual(values = age_group_colors) +
      labs(title = "Wind Effect on Marathon Performance by Age",
           x = "Wind",
           y = "Percentage of Completion Time",
           fill = "Age Group",
           color = "Age Group") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )
```

```r
# SR Effects
SR_sex <- marathon_data %>%
      ggplot(aes(x = as.factor(round(SR, 2)), y = CR_PERCENTAGE, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = Sex), se = FALSE) +
      labs(title = "SR Effect on Marathon Performance by Sex",
           x = "SR",
           y = "Percentage of Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

SR_age <- marathon_data %>%
      ggplot(aes(x = as.factor(round(SR, 2)), y = CR_PERCENTAGE, fill = Age_group, color = Age_group))
      geom_point(alpha = 0.2) +
      geom_smooth(aes(group = Age_group, color = Age_group), se = T) +
      scale_fill_manual(values = age_group_colors) +
      scale_color_manual(values = age_group_colors) +
      labs(title = "SR Effect on Marathon Performance by Age",
           x = "SR",
           y = "Percentage of Completion Time",
           fill = "Age Group",
           color = "Age Group") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

weather_effect <- (RH_sex | RH_age) / (Wind_sex | Wind_age) / (SR_sex | SR_age) +
  plot_layout(guides = "collect", axis_titles = "collect") &
  theme(legend.position = 'bottom') &
  plot_annotation(
    title = "RH, SR and Wind Effect on Marathon Performance",
    theme = theme(
      plot.title = element_text(size = 15, face = "bold")
      )
    )
ggsave("../Plots/weather_effect.png", plot = weather_effect, width = 20, height = 30, dpi = 300)
weather_effect

rh_effect <- (RH_sex | RH_age) +
  plot_layout(guides = "collect", axis_titles = "collect") &
  theme(legend.position = 'bottom') &
  plot_annotation(
    title = "RH Effect on Marathon Performance",
    theme = theme(
      plot.title = element_text(size = 15, face = "bold")
      )
    )
ggsave("../Plots/rh_effect.png", plot = rh_effect, width = 20, height = 30, dpi = 300)
rh_effect
```

```r
wind_effect <- (Wind_sex | Wind_age) +
  plot_layout(guides = "collect", axis_titles = "collect") &
  theme(legend.position = 'bottom') &
  plot_annotation(
    title = "Wind Effect on Marathon Performance",
    theme = theme(
      plot.title = element_text(size = 15, face = "bold")
      )
    )
ggsave("../Plots/wind_effect.png", plot = wind_effect, width = 20, height = 30, dpi = 300)
wind_effect

sr_effect <- (SR_sex | SR_age) +
  plot_layout(guides = "collect", axis_titles = "collect") &
  theme(legend.position = 'bottom') &
  plot_annotation(
    title = "SR Effect on Marathon Performance",
    theme = theme(
      plot.title = element_text(size = 15, face = "bold")
      )
    )
ggsave("../Plots/sr_effect.png", plot = sr_effect, width = 20, height = 30, dpi = 300)
sr_effect

# sex age model
sex_age_model <- glm(CR_PERCENTAGE ~ Sex + Age, data = marathon_data)
kable(summary(sex_age_model)$coefficients, caption = "Sex and Age Model Coefficients")

# RH model
RH_model <- glm(CR_PERCENTAGE ~ RH + Sex + Age + RH * Sex + RH * Age, data = marathon_data)
kable(summary(RH_model)$coefficients, caption = "RH Model Coefficients")

# SR model
SR_model <- glm(CR_PERCENTAGE ~ SR + Sex + Age + SR * Sex + SR * Age, data = marathon_data)
kable(summary(SR_model)$coefficients, caption = "SR Model Coefficients")

# Wind model
Wind_model <- glm(CR_PERCENTAGE ~ Wind + Sex + Age + Wind * Sex + Wind * Age, data = marathon_data)
kable(summary(Wind_model)$coefficients, caption = "Wind Model Coefficients")

# WBGT model
WBGT_model <- glm(CR_PERCENTAGE ~ WBGT + Sex + Age + WBGT * Sex + WBGT * Age, data = marathon_data)
kable(summary(WBGT_model)$coefficients, caption = "WBGT Model Coefficients")

# General model with WBGT
general_model_wbgt <- glm(CR_PERCENTAGE ~ RH + SR + Wind + WBGT + Sex + Age, data = marathon_data)
kable(summary(general_model_wbgt)$coefficients, caption = "Model with WBGT Coefficients")

# General model with flag
general_model_flag <- glm(CR_PERCENTAGE ~ RH + SR + Wind + Flag + Sex + Age, data = marathon_data)
kable(summary(general_model_flag)$coefficients, caption = "Model with Flag Coefficients")
```