Project 1: Exploratory Data Analysis

Morgan Cunningham

2024-10-01

Introduction

This project is a collaborative effort with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. The performance of marathon runners is known to decline as environmental temperatures increase, particularly in long-distance events like a marathon. This performance degradation is especially prominent in older adults, who face challenges in thermoregulation, making it more difficult to dissipate heat. Additionally, there are established differences in endurance performance between men and women, as well as sex-specific physiological responses to temperature and other environmental factors.

The primary objective of this project is to assess the impact of environmental conditions—such as temperature, humidity, solar radiation, and wind speed—on marathon performance across age and gender. This analysis includes a large dataset from several major marathons, spanning performances from age 14 to 85 in both men and women, along with detailed environmental data for each event.

The study will focus on the following aims:

AIM 1: Investigate the effects of increasing age on marathon performance in both men and women.

AIM 2: Explore how various environmental factors influence marathon performance and examine whether these impacts vary across different age groups and genders.

AIM 3: Identify which weather parameters have the largest effect on marathon performance, hypothesizing that increased environmental temperatures and adverse weather conditions negatively impact overall performance.

Project Dataset

The dataset for this project includes marathon race results for men and women and detailed weather data from five major marathon events: Boston, Chicago, New York, Twin Cities (Minneapolis, MN), and Grandma's Marathon (Duluth, MN). These events span from 1993 to 2016, covering 17 to 24 years of performances. Weather data associated with each race includes key environmental parameters such as wet bulb globe temperature, dry bulb temperature, humidity, solar radiation and wind speed.

To prepare for analysis, two data sets including the course records and the aforementioned marathon performance data, were merged. Variable names were standardized for clarity, and race times were converted to seconds to facilitate accurate comparisons. Additionally, an entirely new data set was written with these changes, ensuring the integrity and usability of the data throughout the analysis.

To address missing data values, the table was created below:

Table 1: Missing Data by Race

Number of missing observations per variable stratified by race

	Race Categories				
Variable	B	C	D	NY	TC
Flag Missing	0	0	0	0	0
Dry Bulb Temperature Missing	0	126	116	131	118
Wet Bulb Temperature Missing	0	126	116	131	118
Humidity Missing	0	126	116	131	118
Black Globe Temperature Missing	0	126	116	131	118
Solar Radiation Missing	0	126	116	131	118
Dew Point Missing	0	126	116	131	118
Wind Missing	0	126	116	131	118
Wet Bulb Globe Temperature Missing	0	126	116	131	118

Upon review, we observe that there are missing values across several weather-related variables, including Dry Bulb Temperature, Wet Bulb Temperature, Humidity, Black Globe Temperature, Solar Radiation, Dew Point, Wind, and Wet Bulb Globe Temperature. Each of these variables has 491 missing values, which constitutes approximately 4.25% of the total data set. We see that missing values are not random but tied to specific races and years, suggesting that for certain races in specific years, environmental data may not have been recorded or reported, possibly due to equipment malfunction or inconsistent data collection practices during those events. Since the overall percentage of missing values is below the commonly accepted threshold of 5%, it will have a negligible impact on the results. Therefore, the decision to proceed with observed data only was made.

AIM 1

Investigate the effects of increasing age on marathon performance in both men and women.

We start be creating a summary table of all variables to examine our data set closer.

Variable	**N**	**Overall**, N = 10,440	**Boston**, N = 2,021	**Chicago**, N = 2,256	**Grandma's**, N = 1,717	**New York**, N = 2,711	**Twin Cities**, N = 1,735	**p-value**
Age (years)	10,440	48 (34, 62)	48 (34, 62)	48 (34, 62)	46 (33, 60)	50 (35, 66)	47 (33, 60)	<0.001
Finish Time (s)	10,440	10,919 (9,471, 13,221)	10,547 (9,236, 12,565)	10,750 (9,308, 13,178)	11,184 (9,777, 13,377)	11,107 (9,429, 13,787)	11,068 (9,748, 13,203)	<0.001
Dry Bulb Temp	10,440	12.5 (8.6, 17.3)	9.0 (8.3, 13.8)	14.5 (7.8, 15.7)	18.1 (17.0, 22.0)	12.0 (7.4, 15.1)	11.3 (9.0, 15.7)	<0.001
Wet Bulb Temp	10,440	8.4 (5.4, 13.7)	7.1 (5.4, 8.2)	9.4 (2.9, 12.9)	14.1 (13.7, 16.9)	7.6 (2.9, 11.5)	8.5 (7.3, 11.1)	<0.001
Humidity (%)	10,440	52 (1, 64)	39 (1, 58)	60 (51, 66)	60 (1, 80)	1 (0, 55)	56 (1, 76)	<0.001
Black Globe Temp	10,440	25 (19, 30)	22 (19, 28)	26 (20, 29)	33 (28, 38)	20 (18, 25)	26 (20, 30)	<0.001
Solar Radiation	10,440	513 (368, 608)	708 (574, 800)	479 (439, 536)	731 (520, 833)	417 (309, 546)	481 (348, 545)	<0.001

Variable	N	Overall, N = 10,440	Boston, N = 2,021	Chicago, N = 2,256	Grandma's, N = 1,717	New York, N = 2,711	Twin Cities, N = 1,735	p-value
Dew Point	10,440	5 (0, 11)	3 (0, 6)	6 (-4, 10)	12 (11, 14)	2 (-4, 9)	6 (3, 10)	<0.001
Wind Speed (mph)	10,440	10.0 (7.3, 12.2)	12.0 (8.3, 16.0)	8.4 (5.3, 10.3)	9.0 (7.0, 11.2)	11.0 (9.0, 14.0)	9.2 (6.3, 10.0)	<0.001
Wet Bulb Globe Temp	10,440	12.7 (8.7, 17.7)	10.3 (8.7, 12.7)	13.6 (6.7, 16.4)	18.1 (17.1, 21.8)	10.9 (6.7, 14.1)	12.5 (9.0, 14.4)	<0.001

Table 2 presents a comparative analysis of marathon performance metrics across five race locations: Boston, Chicago, Grandma's, New York, and Twin Cities. The median ages of participants show slight variation, with Grandma's marathon having the youngest participants (median age 46) and Boston having the oldest (median age 48). Finish times also vary, with Grandma's marathon participants generally taking longer to complete the race (median time 11,184 seconds) compared to Boston, where the median finish time is shorter at 10,547 seconds.

Weather conditions show significant differences across the race locations. For instance, Chicago has the highest solar radiation (479) and wet bulb globe temperature (13.6°C), while New York experiences lower solar radiation (417) and a cooler wet bulb globe temperature (10.9°C). Similarly, humidity and wind speed fluctuate between locations, potentially influencing marathon performance. The p-values for all the variables are below 0.001, indicating statistically significant differences between the races in terms of participant characteristics, finish times, and environmental conditions.

We next will look at the effects of age on marathon performance stratified by race, gender, and age.

Figure 1: Effects of Age on Marathon Performance by Race and Gender

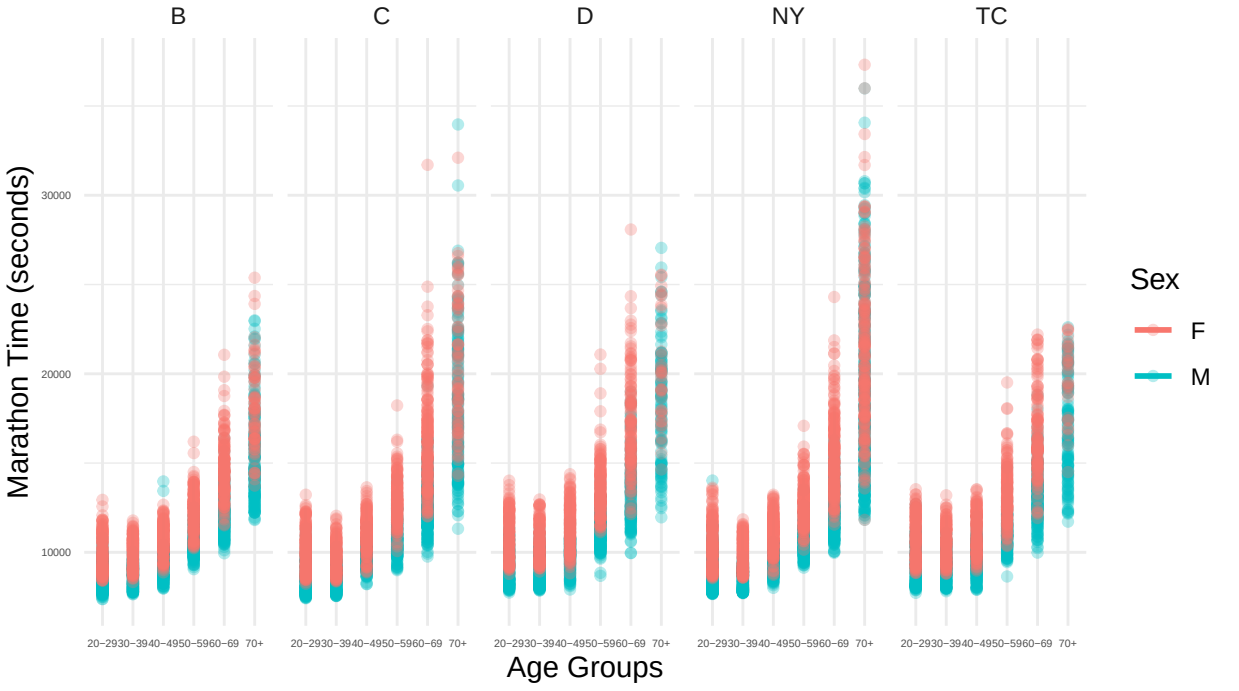


Figure 1 illustrates the effects of age on marathon performance across different race locations, with the

results stratified by gender and grouped into various age brackets. Each point represents individual marathon times in seconds. A general trend is visible where marathon times increase with age across both men and women, especially in the older age groups. Men tend to have slightly faster finish times across most age groups, as indicated by the clustering of blue points (men) lower than the red points (women). The trends show that men tend to outperform women across all age groups, although the difference appears more pronounced in older age brackets. Younger runners, especially those in their 20s and 30s, achieve faster marathon times, indicating that these are likely peak performance years for endurance athletes.

Next we look at the average finish time grouped by age, sex, and race.

Figure 2: Average Marathon Finish Time by Age, Sex, and Race

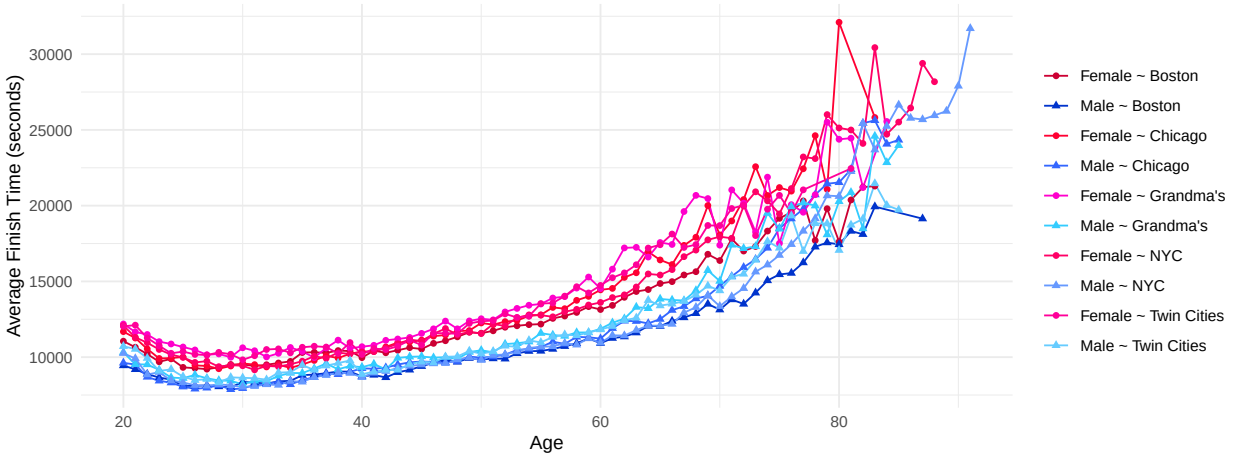


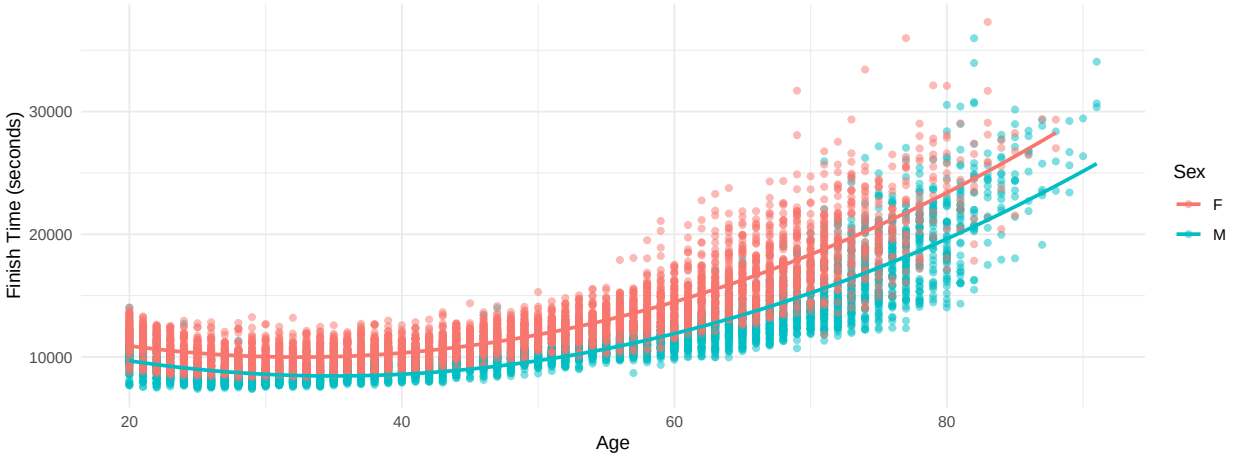
Figure 2 highlights that marathon times are relatively stable between the ages of 20 and 40, with minimal differences across race locations for both men and women. However, after age 40, a clear upward trend in finish times is visible, indicating a steady decline in performance as age increases. Interestingly, Grandma's marathon shows a more pronounced increase in finish times for both men and women, especially after age 60, compared to other races. The variability becomes more extreme after age 70, particularly for women in some races, which suggests increasing performance differences in older age groups.

We look to fit a quadratic polynomial regression model to examine how finish time changes with age for each gender separately because the relationship between age and marathon finish times is not linear and our data shows a curved trend.

```
##
## Call:
## lm(formula = Finish_seconds ~ poly(Age, 2), data = data_men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5846.5  -741.0   -32.2   653.0 15343.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11357.34      20.22  561.63   <2e-16 ***
## poly(Age, 2)1 221021.36    1501.89  147.16   <2e-16 ***
## poly(Age, 2)2 120133.12    1501.89   79.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1502 on 5513 degrees of freedom
## Multiple R-squared:  0.8358, Adjusted R-squared:  0.8357
## F-statistic: 1.403e+04 on 2 and 5513 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Finish_seconds ~ poly(Age, 2), data = data_women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6976.4  -918.5   -85.9   760.5 14239.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12692.62      22.62  561.01   <2e-16 ***
## poly(Age, 2)1 201455.60    1587.60  126.89   <2e-16 ***
## poly(Age, 2)2 104675.44    1587.60   65.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1588 on 4921 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.806
## F-statistic: 1.022e+04 on 2 and 4921 DF,  p-value: < 2.2e-16
```

Figure 3: Marathon Finish Time vs. Age for Men and Women



Both our models and figure 3 reveal that age has a strong non-linear effect on marathon performance for both men and women, with significant linear and quadratic terms. Men's baseline finish time is around 11,357 seconds (3 hours and 9 minutes), while women's is higher at 12,692 seconds (3 hours and 31 minutes). The R-squared values indicate that age explains a large portion of the variation in marathon times for both men (83.6%) and women (80.6%). Figure 3 shows a U-shaped relationship between age and performance, with finish times lowest between ages 20-30, then increasing steadily after 30-40, and accelerating sharply after 50. Women consistently have slower finish times than men, and the performance decline becomes particularly steep after age 40 for both sexes, with greater variability seen in older runners. In summary, increasing age has a clear and substantial effect on marathon performance, with both men and women experiencing a steady decline after their early 30s. Although men tend to have faster finish times than women across all age groups, the rate of performance deterioration due to aging is significant for both sexes, and it becomes especially steep as runners enter their 50s and beyond.

AIM 2

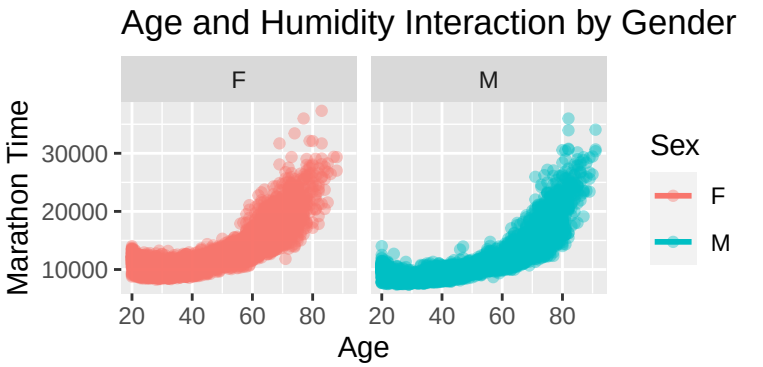
Explore how various environmental factors influence marathon performance and examine whether these impacts vary across different age groups and genders.

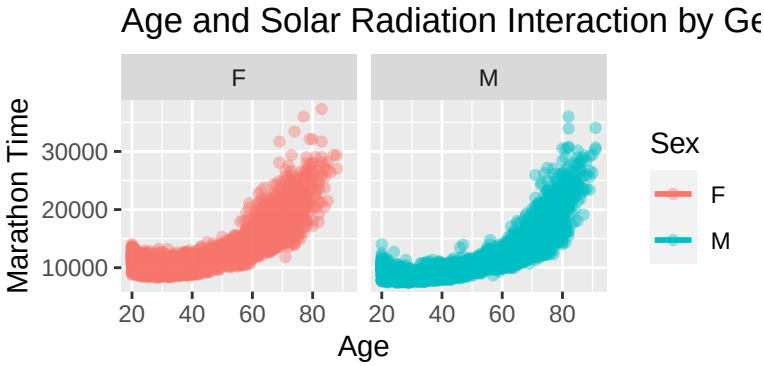
For the second aim, we start by checking for correlations between environmental factors and marathon finish times and then explore these correlations separately for men and women, and across different age groups. We fit a multiple regression model including the environmental factors, age, and gender. We also include interaction terms between sex and environmental factors and age and environmental factors. Age * Sex will test if the effect of age on finish time differs between men and women, Environmental factors * Sex will test whether the influence of environmental conditions on marathon performance is different for men and women, and Environmental factors * Age will check whether the effect of environmental conditions on finish time changes with age.

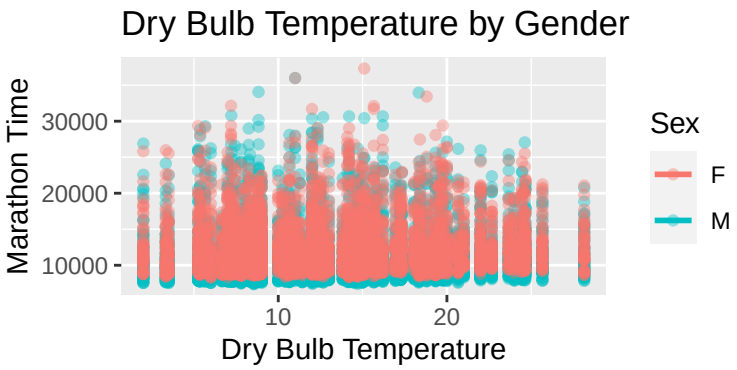
```
##
## Call:
## lm(formula = Finish_seconds ~ Age * Sex + Dry_Bulb_Temp * Sex +
##     Wet_Bulb_Temp * Sex + Humidity * Sex + Black_Globe * Sex +
##     Solar_Radiation * Sex + Dew_Point * Sex + Wind * Sex + Dry_Bulb_Temp *
##     Age + Wet_Bulb_Temp * Age + Humidity * Age + Black_Globe *
##     Age + Solar_Radiation * Age + Dew_Point * Age + Wind * Age,
##     data = final_marathon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6062.6 -1352.0  -491.8   752.5 18790.0
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.513e+03  5.427e+02   2.789  0.00530 **
## Age                   2.355e+02  1.033e+01  22.795  < 2e-16 ***
## SexM                 -1.730e+03  3.701e+02  -4.674 2.98e-06 ***
## Dry_Bulb_Temp        -2.766e+02  9.118e+01  -3.034  0.00242 **
## Wet_Bulb_Temp         5.042e+02  1.942e+02   2.597  0.00942 **
## Humidity              1.471e+01  2.772e+00   5.307 1.13e-07 ***
## Black_Globe          -7.438e+00  2.248e+01  -0.331  0.74077
## Solar_Radiation       4.062e+00  5.438e-01   7.470 8.69e-14 ***
## Dew_Point            -1.527e+02  8.796e+01  -1.736  0.08264 .
## Wind                  1.096e+01  1.966e+01   0.557  0.57734
## Age:SexM             -1.068e+01  2.537e+00  -4.208 2.60e-05 ***
## SexM:Dry_Bulb_Temp   -1.146e+01  5.871e+01  -0.195  0.84517
## SexM:Wet_Bulb_Temp    4.842e+01  1.246e+02   0.388  0.69771
## SexM:Humidity         4.326e+00  1.793e+00   2.414  0.01582 *
## SexM:Black_Globe     -8.702e+00  1.438e+01  -0.605  0.54512
## SexM:Solar_Radiation  4.765e-01  3.503e-01   1.360  0.17374
## SexM:Dew_Point       -3.927e+01  5.639e+01  -0.696  0.48615
## SexM:Wind             7.350e+00  1.269e+01   0.579  0.56242
## Age:Dry_Bulb_Temp     5.131e+00  1.760e+00   2.916  0.00355 **
## Age:Wet_Bulb_Temp    -7.264e+00  3.753e+00  -1.935  0.05298 .
## Age:Humidity         -4.263e-01  5.247e-02  -8.126 4.96e-16 ***
## Age:Black_Globe       3.299e-01  4.464e-01   0.739  0.45997
## Age:Solar_Radiation  -1.065e-01  1.052e-02 -10.120  < 2e-16 ***
## Age:Dew_Point         1.928e+00  1.701e+00   1.134  0.25697
## Age:Wind             -3.288e-01  3.741e-01  -0.879  0.37951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2159 on 10415 degrees of freedom
## Multiple R-squared:  0.6635, Adjusted R-squared:  0.6627
```

```
## F-statistic: 855.7 on 24 and 10415 DF,  p-value: < 2.2e-16
```

Our model reveals significant insights into how age, gender, and environmental factors impact marathon performance. Older runners tend to have slower finish times, with age being a highly significant predictor of performance. Men generally finish marathons faster than women, and this gender difference is statistically significant. Among the weather factors, higher dry bulb temperatures, wet bulb temperatures, humidity, and solar radiation are all associated with slower marathon times, with solar radiation having one of the strongest effects. The model also highlights significant interaction effects. Older runners are more negatively affected by higher humidity and solar radiation, which significantly slows their performance. There are also gender differences, with men and women responding differently to temperature increases, as shown by the significant interaction between sex and dry bulb temperature. Men and women respond differently to weather conditions like dry bulb temperature and humidity. The significance of these interactions (e.g., age with humidity, gender with temperature) indicates that environmental factors disproportionately affect certain groups (e.g., older men, women). However, other weather conditions, such as wet bulb temperature, affect men and women similarly. Overall, the model explains about 66% of the variability in marathon finish times. It shows that age, gender, and environmental conditions all play crucial roles in marathon performance, with older runners and women being more vulnerable to harsh weather conditions like heat and humidity.







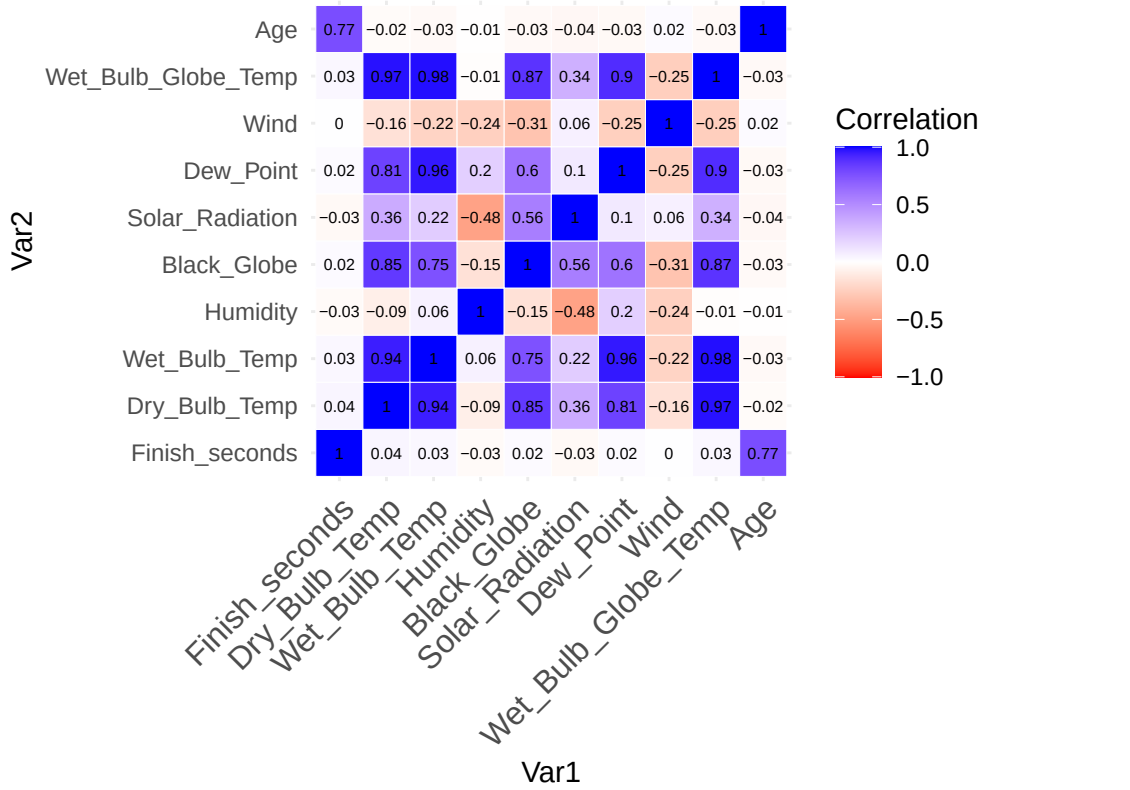
The figures above demonstrate our models for how age and environmental factors (humidity, solar radiation, and temperature) collectively influence marathon performance. The effect of age on marathon performance appears to differ between men and women. Both genders see a performance decline with age, but the interaction plots reveal that certain environmental factors exacerbate this effect, particularly in older age groups.

AIM 3

Identify which weather parameters have the largest effect on marathon performance, hypothesizing that increased environmental temperatures and adverse weather conditions negatively impact overall performance.

To address the hypothesis that increased environmental temperatures and adverse weather conditions negatively impact overall marathon performance, we can use regression modeling and correlation analysis to identify which weather parameters have the largest impact on finish times. Based on the data and previous findings, Dry Bulb Temperature, Wet Bulb Temperature, Wet Bulb Globe Temperature, Humidity, Solar Radiation, Dew Point, and Wind Speed.

Figure 7: Environmental Factors and Marathon Perfor



From the correlation heat map we see there is a weak positive correlation between dry bulb temperature and marathon finish times suggesting that as air temperature increases, marathon performance declines slightly. Wet bulb temperature also shows a weak positive correlation with finish times. This indicates that higher combined temperature and humidity conditions lead to slower performance. WBGT, which a combination of temperature, humidity, and solar radiation, has a weak positive correlation with marathon finish times, meaning more adverse weather conditions (high heat and humidity) slightly slow performance. Interestingly, solar radiation has a weak negative correlation with finish time, suggesting that higher levels of solar radiation could be linked to faster performances, which may be counter intuitive. Humidity shows a weak negative correlation with marathon performance, indicating that higher humidity might actually be linked to slightly better performance, though this effect is minimal and not strong enough to draw concrete conclusions from. The weather parameters that appear to have the largest (though still weak) effects on marathon performance are Dry Bulb Temperature, Wet Bulb Temperature, and WBGT, all of which support the hypothesis that higher temperatures and more adverse weather conditions negatively impact marathon performance.

Next we perform multiple regression analysis in order to confirm statistical significance on which weather factors have the largest effect on performance when controlling for other variables like age and gender.

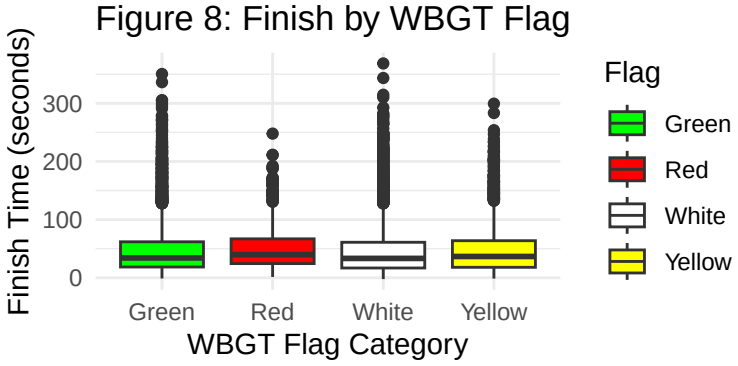
```
##
## Call:
## lm(formula = Finish_seconds ~ Age + Sex + Dry_Bulb_Temp + Wet_Bulb_Temp +
##     Humidity + Black_Globe + Solar_Radiation + Dew_Point + Wind,
##     data = final_marathon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5249.3 -1384.9  -514.6   750.0 19023.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4397.4421   186.1231  23.627  < 2e-16 ***
## Age               173.1326     1.2641 136.963  < 2e-16 ***
## SexM            -1871.6510    42.8828 -43.646  < 2e-16 ***
## Dry_Bulb_Temp     -28.4742    29.3812  -0.969  0.33250
## Wet_Bulb_Temp     163.8416    62.4084   2.625  0.00867 **
## Humidity           -3.9195     0.8990  -4.360 1.31e-05 ***
## Black_Globe         2.8188     7.1833   0.392  0.69476
## Solar_Radiation    -0.8243     0.1753  -4.702 2.61e-06 ***
## Dew_Point         -72.5440    28.2407  -2.569  0.01022 *
## Wind               -1.8666     6.3566  -0.294  0.76904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2178 on 10430 degrees of freedom
## Multiple R-squared:  0.6571, Adjusted R-squared:  0.6568
## F-statistic:  2221 on 9 and 10430 DF,  p-value: < 2.2e-16
```

The most impactful weather-related variables are Wet Bulb Temperature, Humidity, Solar Radiation, and Dew Point. Wet Bulb Temperature has the strongest positive effect, suggesting that hot and humid conditions significantly slow down marathon performance. Solar Radiation surprisingly has a negative effect, indicating that higher solar radiation is associated with slightly faster finish times, which could be influenced by other favorable conditions on sunny race days. Dew Point also shows a significant negative relationship.

Next, we explore the Flag parameter which categorizes Wet Bulb Globe Temperature (WBGT) based on heat illness risk levels. We want to show how marathon performance is impacted by the different flag levels below:

- White: WBGT < 10°C
- Green: WBGT 10-18°C
- Yellow: WBGT 18-23°C
- Red: WBGT 23-28°C
- Black: WBGT > 28°C



We see from figure 8 that runners in the Green flag category have the most consistent and favorable finish times, suggesting optimal running conditions. In contrast, Red flag conditions show a slightly higher median finish time, though the distribution is less variable, indicating moderately challenging conditions. The White flag and Yellow flag categories exhibit more variability and numerous outliers, suggesting that some runners struggle in these conditions. Overall, as the WBGT flag moves from more favorable (Green) to more challenging (Red/Yellow), there is an observable trend towards longer and more varied finish times, highlighting the impact of environmental conditions on marathon performance. Next we use an ANOVA test to assess whether the differences between the WBGT flag categories are actually statistically significant.

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## Flag            3    17841    5947   2.875 0.0348 *
## Residuals   10436 21586162    2068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than 0.05, we can conclude that marathon finish times vary significantly across the different WBGT flag categories. To determine which specific flag categories differ from one another, we can run a Tukey's HSD test. This will allow us to perform pairwise comparisons between the flag categories to identify where the significant differences lie.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Finish ~ Flag, data = final_marathon)
##
## $Flag
##                     diff         lwr         upr     p adj
## Red-Green      4.7114239  -0.5847787 10.00762642 0.1013821
## White-Green   -0.6027354  -3.2295674  2.02409667 0.9352698
## Yellow-Green   1.7398631  -1.4728482  4.95257439 0.5046833
## White-Red     -5.3141592 -10.6789183  0.05059982 0.0533030
## Yellow-Red    -2.9715608  -8.6462331  2.70311164 0.5338535
## Yellow-White   2.3425985  -0.9819153  5.66711223 0.2683138
```

Although the overall ANOVA showed significant differences in finish times across the WBGT flag categories, the pairwise comparisons do not show any specific flag categories having significant differences after adjusting for multiple comparisons. The most notable result is the White vs Red comparison, which is close to significance and suggests that finish times might be slower in Red flag conditions (higher WBGT) compared to White (lower WBGT).

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
# set wd, create DFs, load in proper packages
setwd("/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1")
course_record <- read.csv("course_record.csv")
marathon_dates <- read.csv("marathon_dates.csv")
marathon_df <- read.csv("project1.csv")
aqi_df <- read.csv("aqi_values.csv")
library(gtsummary)
library(dplyr)
```

```r
library(gt)
library(tidyr)
library(tidyverse)
library(HDSinRdata)
library(ggplot2)
library(readr)
library(reshape2)
library(car)
# change column names (marathon_df)
marathon_df <- marathon_df %>%
  rename(Race = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
         Sex = Sex..0.F..1.M.,
         Age = Age..yr.,
         Finish = X.CR,
         Dry_Bulb_Temp = Td..C,
         Wet_Bulb_Temp = Tw..C,
         Humidity = X.rh,
         Black_Globe = Tg..C,
         Solar_Radiation = SR.W.m2,
         Dew_Point = DP,
         Wet_Bulb_Globe_Temp = WBGT
         )
# change race/sex variable names
marathon_df <- marathon_df %>%
  mutate(Race = case_when(Race == "0" ~ "B",
                          Race == "1" ~ "C",
                          Race == "2" ~ "NY",
                          Race == "3" ~ "TC",
                          Race == "4" ~ "D")) %>%
  mutate(Sex = case_when(Sex == "0" ~ "F",
                         Sex == "1" ~ "M"))

course_record <- course_record %>%
  rename(Sex = Gender)
# join marathon_df and course_record
marathon_time <- left_join(marathon_df, course_record,
                           by=c("Race", "Sex", "Year"))
# write new CSV for marathon_time
write_csv(marathon_time, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1/ma
marathon_time <- read.csv("marathon_time.csv")
# create actual time column
# convert "HH:MM:SS" to seconds
time_to_seconds <- function(time_str) {
  parts <- as.numeric(unlist(strsplit(time_str, ":")))
  return(parts[1] * 3600 + parts[2] * 60 + parts[3])
}

# apply to convert each CR to seconds
marathon_time$CR_seconds <- sapply(marathon_time$CR, time_to_seconds)

# calculate actual race time in seconds based on Finish percentage off the CR time
marathon_time$Actual_Time_Seconds <- marathon_time$CR_seconds * (1 + (marathon_time$Finish / 100))
marathon_df <- marathon_time %>%
```

```r
  rename(Finish_seconds = Actual_Time_Seconds)
# write new CSV for marathon_df with actual course times
write_csv(marathon_df, "/Users/morgancunningham/Desktop/PHP 2550 Practical Data Analysis/Project 1/mara
# missing data by Race
missing_tb2 <- marathon_df %>%
  group_by(Race) %>%
  summarise(
    Flag_Missing = sum(is.na(Flag)),
    Dry_Bulb_Temp_Missing = sum(is.na(Dry_Bulb_Temp)),
    Wet_Bulb_Temp_Missing = sum(is.na(Wet_Bulb_Temp)),
    Humidity_Missing = sum(is.na(Humidity)),
    Black_Globe_Missing = sum(is.na(Black_Globe)),
    Solar_Radiation_Missing = sum(is.na(Solar_Radiation)),
    Dew_Point_Missing = sum(is.na(Dew_Point)),
    Wind_Missing = sum(is.na(Wind)),
    Wet_Bulb_Globe_Temp_Missing = sum(is.na(Wet_Bulb_Globe_Temp))
  )

missing_tb2 <- as.data.frame(t(missing_tb2))
colnames(missing_tb2) <- missing_tb2[1, ]
missing_tb2 <- missing_tb2[-1, ]
missing_tb2$Variable <- factor(row.names(missing_tb2))
missing_tb2 <- missing_tb2[, c(ncol(missing_tb2), 1:(ncol(missing_tb2) - 1))]

missing_tb2 <- missing_tb2 %>%
  mutate(Variable = case_when(
    Variable == "Flag_Missing" ~ "Flag Missing",
    Variable == "Dry_Bulb_Temp_Missing" ~ "Dry Bulb Temperature Missing",
    Variable == "Wet_Bulb_Temp_Missing" ~ "Wet Bulb Temperature Missing",
    Variable == "Humidity_Missing" ~ "Humidity Missing",
    Variable == "Black_Globe_Missing" ~ "Black Globe Temperature Missing",
    Variable == "Solar_Radiation_Missing" ~ "Solar Radiation Missing",
    Variable == "Dew_Point_Missing" ~ "Dew Point Missing",
    Variable == "Wind_Missing" ~ "Wind Missing",
    Variable == "Wet_Bulb_Globe_Temp_Missing" ~ "Wet Bulb Globe Temperature Missing"
  ))

gt_table <- gt(missing_tb2) %>%
  tab_header(
    title = "Table 1: Missing Data by Race",
    subtitle = "Number of missing observations per variable stratified by race"
  ) %>%
  tab_spanner(
    label = "Race Categories",
    columns = colnames(missing_tb2)[-1]
  )

gt_table
# total percent of missing values in the dataset
total_missing <- sum(is.na(marathon_df))
total_values <- prod(dim(marathon_df))
percentage_missing <- (total_missing / total_values) * 100
```

```r
# percentage of missing data is negligible since <5%, Use observed data only
# y-axis average finish, x-axis age, color F/M
#group by into age groups, Under 19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+
# Under 19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70+, typical age grou
final_marathon <- marathon_df %>%
  mutate(
    age_group = case_when(
      Age <= 19 ~ "19 and Under",
      Age > 19 & Age <= 29 ~ "20-29",
      Age > 29 & Age <= 39 ~ "30-39",
      Age > 39 & Age <= 49 ~ "40-49",
      Age > 49 & Age <= 59 ~ "50-59",
      Age > 59 & Age <= 69 ~ "60-69",
      Age > 69 ~ "70+"
    ),
    age_group = factor(
      age_group,
      level = c("Under 19", "20-29","30-39", "40-49", "50-59", "60-69", "70+")
    )
  )

 final_marathon <- na.omit(final_marathon)
# TABLE 1
final_marathon %>%
  mutate(Sex = ifelse(Sex == "M", "Male", "Female"),
         Race = case_when(Race == "B" ~ "Boston",
                          Race == "C" ~ "Chicago",
                          Race == "NY" ~ "New York",
                          Race == "TC" ~ "Twin Cities",
                          Race == "D" ~ "Grandma's")) %>%
    tbl_summary(
    include = c(Age, Finish_seconds, Dry_Bulb_Temp, Wet_Bulb_Temp, Humidity, Black_Globe,
                Solar_Radiation, Dew_Point, Wind, Wet_Bulb_Globe_Temp),
    by = Race,
    label = list(
      Age = "Age (years)",
      Finish_seconds = "Finish Time (s)",
      Dry_Bulb_Temp = "Dry Bulb Temp",
      Wet_Bulb_Temp = "Wet Bulb Temp",
      Humidity = "Humidity (%)",
      Black_Globe = "Black Globe Temp",
      Solar_Radiation = "Solar Radiation",
      Dew_Point = "Dew Point",
      Wind = "Wind Speed (mph)",
      Wet_Bulb_Globe_Temp = "Wet Bulb Globe Temp"
    ),
    missing = "no"
  ) %>%
  add_n() %>%
  add_p() %>%
  add_overall() %>%
  bold_labels() %>%
  modify_header(label = "**Variable**") %>%
```

```r
  modify_spanning_header(all_stat_cols() ~ "**Marathon Race**")
ggplot(final_marathon, aes(x = age_group, y = Finish_seconds, color = Sex)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method=lm, se = FALSE) +
  labs( title = "Figure 1: Effects of Age on Marathon Performance by Race and Gender",
        x = "Age Groups",
        y = "Marathon Time (seconds)") +
  facet_grid(~Race) +
  theme_minimal()+
  theme(axis.text = element_text(size = 4))
# average finish time grouped by Age, Sex, and Race
avg_finish_time <- aggregate(Finish_seconds ~ Age + Sex + Race, data = final_marathon, FUN = mean)

warm_colors <- c("#CC0033", "#FF0033", "#FF0066", "#FF0099", "#FF00CC")  # For females
cold_colors <- c("#0033CC", "#3366FF", "#6699FF", "#66CCFF", "#33CCFF")  # For males

races <- unique(final_marathon$Race)

color_mapping <- setNames(
  c(warm_colors, cold_colors),
  c(paste0("F.", races), paste0("M.", races))
)

custom_labels <- c(
  "Female ~ Boston", "Male ~ Boston", "Female ~ Chicago", "Male ~ Chicago",
  "Female ~ Grandma's", "Male ~ Grandma's", "Female ~ NYC", "Male ~ NYC",
  "Female ~ Twin Cities", "Male ~ Twin Cities"
)

ggplot(avg_finish_time, aes(x = Age, y = Finish_seconds, color = interaction(Sex, Race), shape = intera
  geom_line() +
  geom_point() +
  scale_color_manual(values = color_mapping, labels = custom_labels) +
  scale_shape_manual(values = c(16, 17, 16, 17, 16, 17, 16, 17, 16, 17), labels = custom_labels) +
  labs(title = 'Figure 2: Average Marathon Finish Time by Age, Sex, and Race',
       x = 'Age',
       y = 'Average Finish Time (seconds)') +
  theme_minimal() +
  theme(legend.title = element_blank(), legend.position = "right")

data_men <- final_marathon %>% filter(Sex == "M")
# quadratic polynomial model for men
model_men <- lm(Finish_seconds ~ poly(Age, 2), data = data_men)
summary(model_men)

data_women <- final_marathon %>% filter(Sex == "F")
# quadratic polynomial model for women
model_women <- lm(Finish_seconds ~ poly(Age, 2), data = data_women)
summary(model_women)


# Plot data for both men and women
ggplot(final_marathon, aes(x = Age, y = Finish_seconds, color = Sex)) +
```

```r
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
  labs(title = "Figure 3: Marathon Finish Time vs. Age for Men and Women",
       x = "Age", y = "Finish Time (seconds)") +
  theme_minimal()
# multiple regression model
model <- lm(Finish_seconds ~ Age * Sex + Dry_Bulb_Temp * Sex + Wet_Bulb_Temp * Sex +
              Humidity * Sex + Black_Globe * Sex + Solar_Radiation * Sex + Dew_Point * Sex + Wind * Sex
              Dry_Bulb_Temp * Age + Wet_Bulb_Temp * Age + Humidity * Age +
              Black_Globe * Age + Solar_Radiation * Age + Dew_Point * Age + Wind * Age,
            data = final_marathon)

summary(model)
par(mfrow=c(1, 3))
# Age and Humidity, stratified by Gender
ggplot(final_marathon, aes(x = Age, y = Finish_seconds, color = Sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", formula = y ~ poly(Age, 2) + Humidity, se = FALSE) +
  labs(title = "Age and Humidity Interaction by Gender",
       x = "Age", y = "Marathon Time") +
  facet_wrap(~ Sex)

# Age and Solar Radiation, stratified by Gender
ggplot(final_marathon, aes(x = Age, y = Finish_seconds, color = Sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", formula = y ~ poly(Age, 2) + Solar_Radiation, se = FALSE) +
  labs(title = "Age and Solar Radiation Interaction by Gender",
       x = "Age", y = "Marathon Time") +
  facet_wrap(~ Sex)

# Dry Bulb Temperature and Gender
ggplot(final_marathon, aes(x = Dry_Bulb_Temp, y = Finish_seconds, color = Sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", formula = y ~ Dry_Bulb_Temp * Sex, se = FALSE) +
  labs(title = "Dry Bulb Temperature by Gender",
       x = "Dry Bulb Temperature", y = "Marathon Time")

# correlation matrix for environmental factors and finish time
cor_matrix <- final_marathon %>%
  select(Finish_seconds, Dry_Bulb_Temp, Wet_Bulb_Temp, Humidity, Black_Globe, Solar_Radiation, Dew_Poin
  cor()

melted_cor <- melt(cor_matrix)

# correlation heatmap
ggplot(data = melted_cor, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white", midpoint = 0, limit = c(-1, 1), space
                       name="Correlation") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 2) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1)) +
  coord_fixed() +
```

```r
  labs(title = "Figure 7: Environmental Factors and Marathon Performance")

# multiple regression model including age, gender, and weather
model_weather <- lm(Finish_seconds ~ Age + Sex + Dry_Bulb_Temp + Wet_Bulb_Temp +
                      Humidity + Black_Globe + Solar_Radiation + Dew_Point + Wind,
                    data = final_marathon)

summary(model_weather)
# Finish Times by WBGT Flag Category
ggplot(final_marathon, aes(x = Flag, y = Finish, fill = Flag)) +
  geom_boxplot() +
  labs(title = "Figure 8: Finish by WBGT Flag",
       x = "WBGT Flag Category", y = "Finish Time (seconds)") +
  theme_minimal() +
  scale_fill_manual(values = c("White" = "white", "Green" = "green", "Yellow" = "yellow",
                               "Red" = "red", "Black" = "black"))

# ANOVA test
anova_result <- aov(Finish ~ Flag, data = final_marathon)
summary(anova_result)
# Tukey's HSD test
tukey_result <- TukeyHSD(anova_result)
print(tukey_result)
```