Exploratory Analysis on the Effects of Weather Conditions and Aging on Marathon Performance

Rachel Yost

Abstract

Previous studies have found that environmental temperature decreases marathon performance. Additionally, older adults experience difficulties with thermoregulation. Using data provided by Dr. Brett Romano and Dr. Matthew Ely from the Department of Health Sciences at Providence College, we performed an exploratory data analysis aiming to examine the effects of increasing age on marathon performance in men and women; explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender; and identify the weather parameters that have the largest impact on marathon performance. We hypothesized that the weather would have a stronger negative impact on older runners, and that effects would be similar between men and women. Furthermore, we suspected that relative humidity, temperature, and solar radiation would have the strongest impact on runners. Based on the results of my exploratory data analysis, weather has a stronger impact on runners as age increases, but effects do not vary significantly by gender. Furthermore, based on a linear regression and correlations between race results and weather conditions, we found that relative humidity, air quality, and solar radiation have the largest impact on marathon performance. Future research is needed to statistically analyze the impact of weather, age, and gender on marathon performance.

Introduction

In recent years, we've seen an increase in marathon participants (Reusser et al., 2021), but also an increase in heatwaves (NOAA, 2024), which could impact athletic performance (Ely et al., 2007). Also, thermoregulation ability is known to decrease with age, affecting sweat gland function, skin blood flow, cardiac output, and blood flow redistribution (Kenny and Munce, 2023). This could potentially cause problems for older runners who are racing in unfavorable weather conditions. This has been shown in the New York marathon, where older runners were more affected by high temperatures and humidity (Knechtle et al. 2021). Differences in marathon performance are also seen across genders. For example, men are more likely to slow down over the course of the marathon (Deanor et al. 2015).

Using data provided by Dr. Brett Romano and Dr. Matthew Ely from the Department of Health Sciences at Providence College, we performed an exploratory data analysis with the following three aims: **1. Examine effects of increasing age on marathon performance in men and women. 2. Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender. 3. Identify the weather parameters that have the largest impact on marathon performance.** We hypothesized that aging would decrease marathon performance, and that weather would have a stronger negative effect on marathon performance as age increased, but effects would be similar between men and women. Additionally, we hypothesized that humidity, solar radiation, and temperature would all have a strong impact on marathon performance, and that wind and air quality would have a lesser impact.

Methods

Data Description

The provided data from the Department of Health Sciences at Providence College contained results from the Boston, Chicago, New York, Twin Cities, and Grandma's marathons across 15-20 years. Each row of the data contained the top single-age performances from males and females between the ages of 14 and 85 for each of the 5 marathons. The following weather parameters were also included: dry bulb temperature, wet bulb temperature (which factors in humidity), percent relative humidity, black globe temperature (factors in solar radiation), wind speed in km/hr, dew point, and WBGT (calculated as (.07 * Wet-bulb temperature) + (.02 * black globe temperature) + (0.1 * dry bulb temperature)). Additionally, a variable describing the flag color used to warn participants of weather conditions was used. The flag color was determined based on the following conditions: White = WBGT $<$ 10C, Green = WBGT between 10-18C, Yellow = WBGT between 18-23C, Red = WBGT between 23-28C, and Black = WBGT $>$ 28C.

Instead of the finishing time for each participant, the data provided the percent off the current course record for each gender. This value was used for most analyses to allow for more accurate comparisons across races, which may have varying difficulty.

Additionally, we included air quality index measurements from the RAQSAPI package (Mccrowey et al. 2023), which retrieves air quality data from the United States Environmental Protection Agency. From this package, we selected to use air quality index (aqi) measurements from core based statistical area (CBSA) codes 14460 (Boston-Cambridge-Newton), 35620 (New York-Newark-Jersey City), 33460 (Minneapolis-St. Paul-Bloomington), 16980 (Chicago-Naperville-Elgin), and 20260 (Duluth) , from the dates of each of the marathons over the time period from 1993-2016. Each of these codes is a marathon location. We chose to use the values from the 24 hour sample duration, which is recorded in micrograms/cubic meter. The data contained some duplicates, so we removed duplicate values. Since each CBSA code contained multiple sites where aqi was measured, we averaged the data between all sites for each CBSA code for each date.

Missing data and data preprocessing

My first step in preprocessing the data was to rename the columns so that they were easier to handle within the code. We removed spaces and symbols from variable names. Next, we checked the data types of each variable to make sure they made sense for my analyses. We changed "Race", "Year", "Sex", and "Flag" from character/numeric to factors. We also changed the values for Race, so that each observation was the name of the marathon rather than a numerical code. We did the same for Sex, and re-coded 0 as "Female" and 1 as "Male".

I also calculated the actual time of each participant in minutes, using the course records for each race for each year by gender. We converted the course record from hours into minutes, and then performed the following calculation: Time in minutes = ((Percent off course record)/100) + 1) * Course Record in minutes).

Based on data found from the analyses shown in Figure 1., we also created a variable that groups participants by age. Runners younger than 20 were in "Young", ages 20-49 were in "Middle", 50-64 were in "Older", and runners 65 or older were in "Senior". These groups were made based on trends in performance as age increased.

Next,we checked for missing data and found that weather variables, not including aqi, were missing for the Twin Cities, Chicago, and New York marathons in 2011, and the Grandma's marathon in 2012. AQI data was missing for 32 of the races.

I checked the distribution of each variable, and we noticed that 32% of the values for relative humidity were less than 1, but then the other 68% of the values were all greater than 33. This seemed suspicious, and after comparing the relative humidities that were less than 1 with several historical weather reports from those

dates, we decided that the data was just recorded in a different format, and that any values less than 1 could be multiplied by 100 to get the percent relative humidity.

Results

The data included 11,564 runners, of which 47% were female and 53% were male. Most races had a green flag (42%), while few had a red flag (5.3%). None of the data included the most severe flag option, black. The age groups were binned so that around 5% of participants were in "Young", 50% of participants were in "Middle", 25% in "Older", and 20% in "Senior". These statistics and other characteristics can be seen in Table 1 and Table 2. For continuous variables, medians and the first and third quartile were calculated. For categorical variables, counts and percentages are shown.

Table 1: Race Characteristics

Characteristic	**Boston** N = 18	**Chicago** N = 21	**Grandma's** N = 17	**New York** N = 23	**Twin Cities** N = 17
Flag					
White	9 (50%)	6 (30%)	0 (0%)	11 (50%)	5 (31%)
Green	7 (39%)	12 (60%)	6 (38%)	7 (32%)	7 (44%)
Yellow	1 (5.6%)	1 (5.0%)	8 (50%)	4 (18%)	3 (19%)
Red	1 (5.6%)	1 (5.0%)	2 (13%)	0 (0%)	1 (6.3%)
Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Dry Bulb	9.0 (8.3, 13.8)	13.6 (7.4, 15.3)	18.7 (16.3, 22.0)	11.5 (7.4, 15.1)	11.6 (9.0, 16.5)
Wet Bulb	7.2 (5.4, 8.2)	9.3 (2.7, 12.5)	14.3 (13.7, 16.2)	7.2 (2.9, 11.5)	8.7 (6.6, 12.6)
% Relative Humidity	37 (1, 58)	60 (53, 68)	58 (1, 78)	1 (0, 55)	53 (1, 70)
Black Globe	23 (19, 28)	26 (21, 29)	34 (28, 38)	20 (18, 25)	26 (20, 30)
Solar Radiation	721 (574, 800)	470 (437, 518)	736 (571, 838)	393 (309, 546)	488 (355, 541)
Dew Point	3 (0, 6)	6 (-2, 10)	12 (11, 14)	2 (-4, 9)	6 (3, 10)
Wind	11.8 (8.3, 16.0)	8.0 (5.3, 10.2)	9.3 (7.7, 11.2)	11.2 (9.0, 14.0)	9.3 (6.5, 10.0)
WBGT	9.9 (8.7, 12.7)	13.1 (7.2, 16.1)	18.1 (16.0, 21.0)	10.2 (6.7, 14.1)	12.6 (9.0, 16.3)
AQI	44 (30, 55)	52 (42, 69)	22 (14, 27)	47 (28, 54)	31 (19, 43)

[1] n (%); Median (Q1, Q3)

Table 2: Participant Characteristics

Characteristic	**Boston** N = 2,085	**Chicago** N = 2,549	**Grandma's** N = 1,994	**New York** N = 2,925	**Twin Cities** N = 1,990
Sex					
Female	981 (47%)	1,209 (47%)	933 (47%)	1,400 (48%)	922 (46%)
Male	1,104 (53%)	1,340 (53%)	1,061 (53%)	1,525 (52%)	1,068 (54%)
Age	47 (32, 61)	46 (30, 61)	44 (29, 58)	49 (33, 65)	44 (30, 59)
% Course Record	32 (18, 56)	38 (20, 67)	38 (20, 62)	37 (19, 69)	36 (19, 63)
Time	176 (155, 208)	182 (157, 222)	190 (165, 227)	186 (158, 230)	188 (164, 226)
Age Binned					
Young	67 (3.2%)	179 (7.0%)	176 (8.8%)	92 (3.1%)	149 (7.5%)
Middle	1,078 (52%)	1,257 (49%)	1,016 (51%)	1,378 (47%)	1,017 (51%)
Older	532 (26%)	629 (25%)	505 (25%)	687 (23%)	506 (25%)
Senior	408 (20%)	484 (19%)	297 (15%)	768 (26%)	318 (16%)

[1] n (%); Median (Q1, Q3)

To start, we plotted percent off of course record by age for each of the five marathons. As shown in Figure 1, the youngest runners gradually improved in performance until a peak between ages 27-32. Males and females had slightly different ages of peak performance by race. Males and females both decreased in performance steadily until their late 50s-early 60s, and then performance began decreasing quickly. Running performance in females generally declined more with age, compared to males. However, performance became similar after about age 75. This suggests a possible interaction between age and gender on marathon performance, with women beginning to quickly decrease in pace at an earlier age than men. These results were consistent across races.

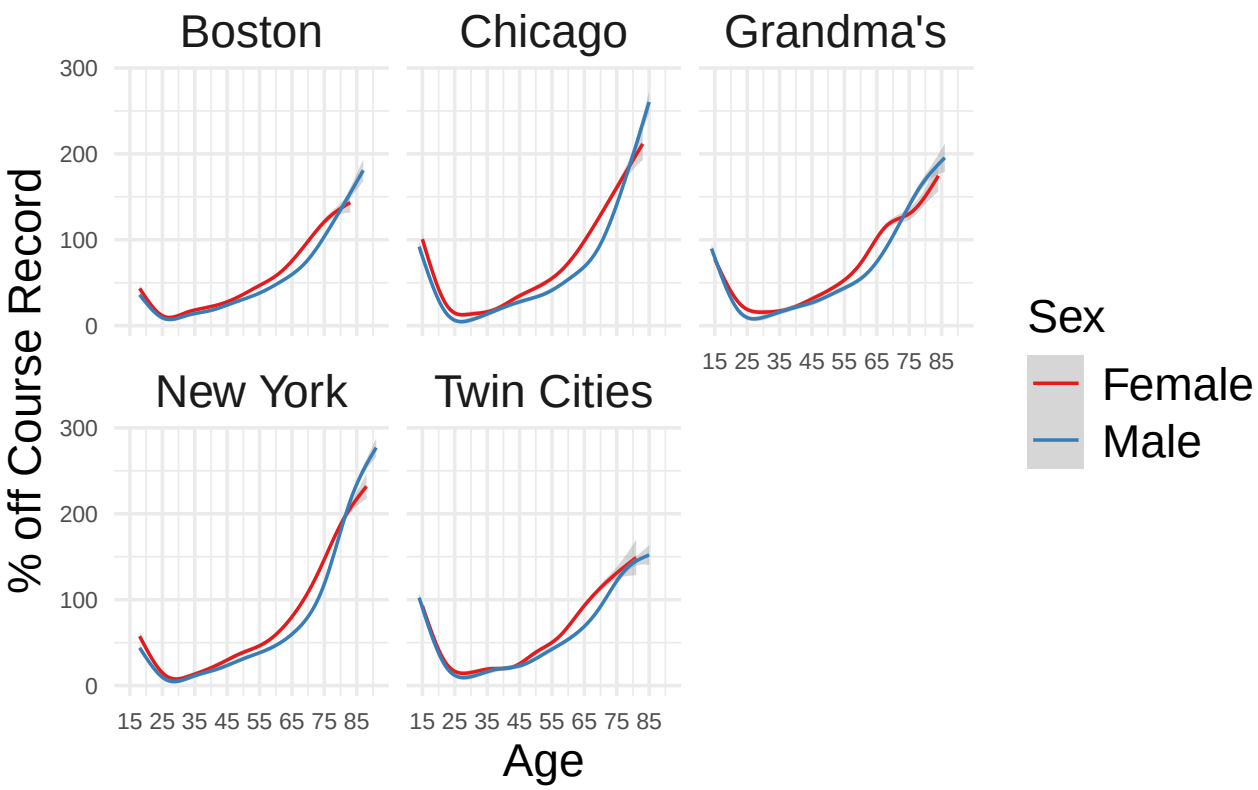

Figure 1: % Off of Course Record by Age, Race, and Gender

Next, we examined what ages were actually winning the marathons. We found the best performance for each race and year for each sex, and found that the most female winners were 28, and most male winners were 29 (Figure 2). The mean age of the female winners was 30.31, and the mean age of male winners was 28.99.

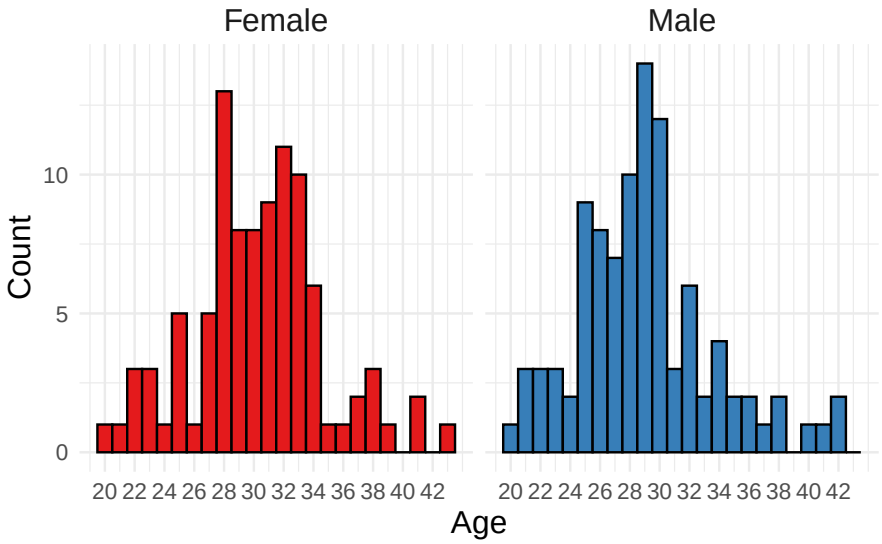

Figure 2: Histogram of Ages of Marathon Winners by Gender

I also visualized how average marathon times improved over time over the years, faceted by gender. As we can see from Figure 3, on average, both men and women have been improving over time since 1993. Women saw a quick increase in pace in the 1990s, and then plateaued from 2000-2008, but gradually improved from 2008-2016 (Figure 3). Men have been steadily getting faster, aside from a plateau between 2000 and 2007 (Figure 3).

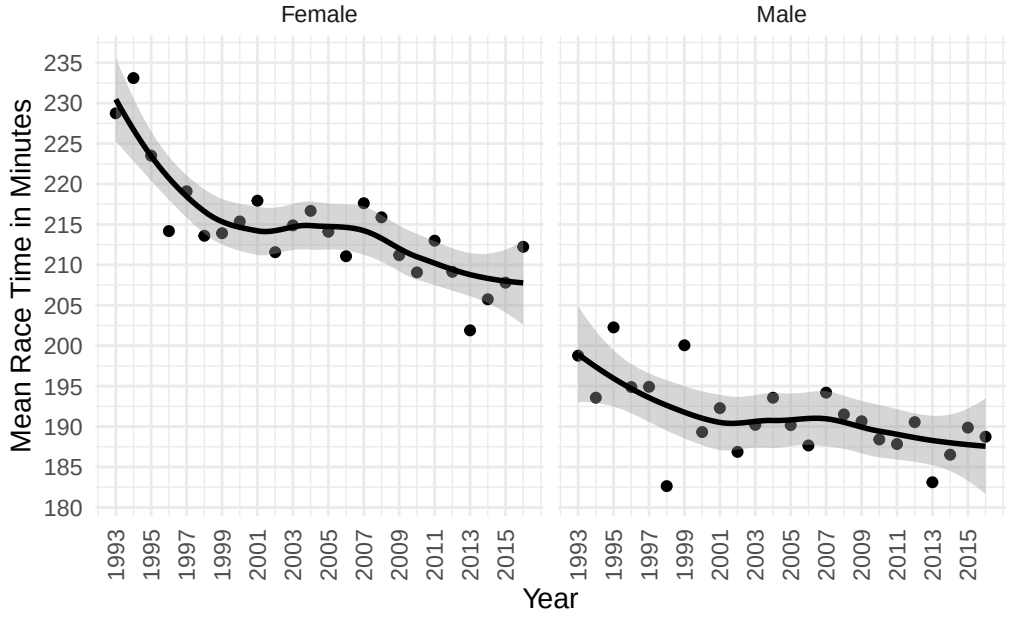

Figure 3: Marathon Time in Minutes by Year

To determine if a red flag results in a weaker average performance in each age group, we plotted the mean percent off course record for each of the binned age groups (Young, Middle, Older, and Senior; previously described in *Methods*), for each flag for each race (Figure 4). Error bars represent 95% confidence intervals based on standard error. The only races that showed differences in performance based on flag were the Boston, Chicago, and Twin Cities marathons. In the Boston Marathon, the mean percent off course record increases between the white and red flags for all age groups except for "Young" and "Older", and this increase appears to be largest in the "Senior" group. For the Twin Cities marathon, mean percent off course record increases for the "Middle" and "Older" age groups between the white flag and red flag. For the Chicago marathon, mean percent off course record also only increases between the white and red flags for the "Middle" and "Older" groups only. However, there were few (16) results recorded in the "Senior" age group with a red flag, so the lack of significance may be due to the small sample size.

The difference in performance by age group may not necessarily be due to physiological effects caused by the weather. For example, runners could be aware that the flag was red and purposefully slowed down for their own well being and comfort, rather than being physically unable to go any faster. Older runners may believe that they are more susceptible to health problems from running in the heat than younger runners, so they choose to just run the race at a slower than usual pace.

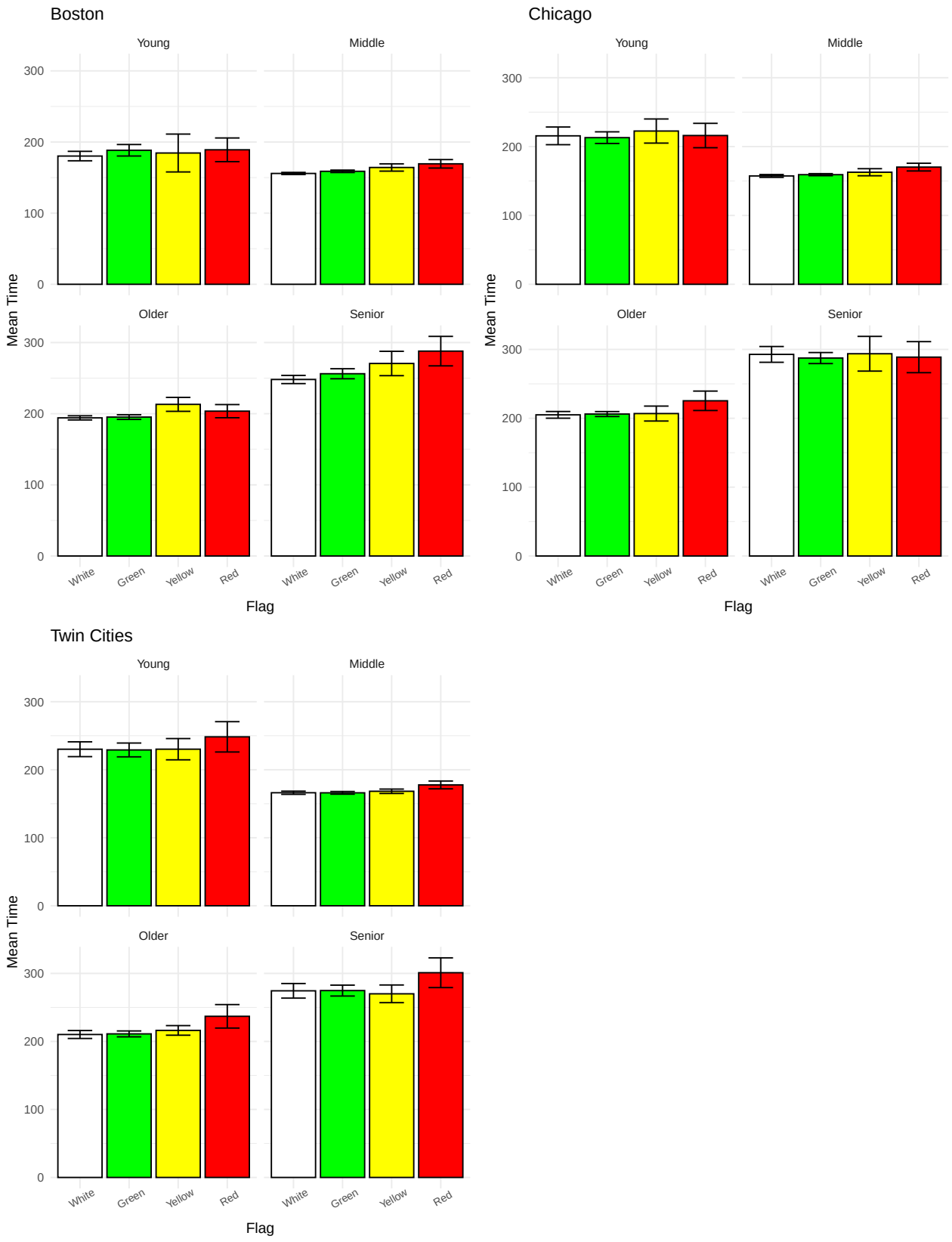

Figure 4: Mean Time by Age and Flag

To see how the weather variables are correlated with each other, we produced a plot of the spearman correlations between the variables. Percent off of course record appears to only be strongly correlated with age. The weather variables are correlated with each other, which is to be expected since many of the variables are calculated based on other variables. Humidity and solar radiation are negatively correlated, and wind and aqi are negatively correlated.

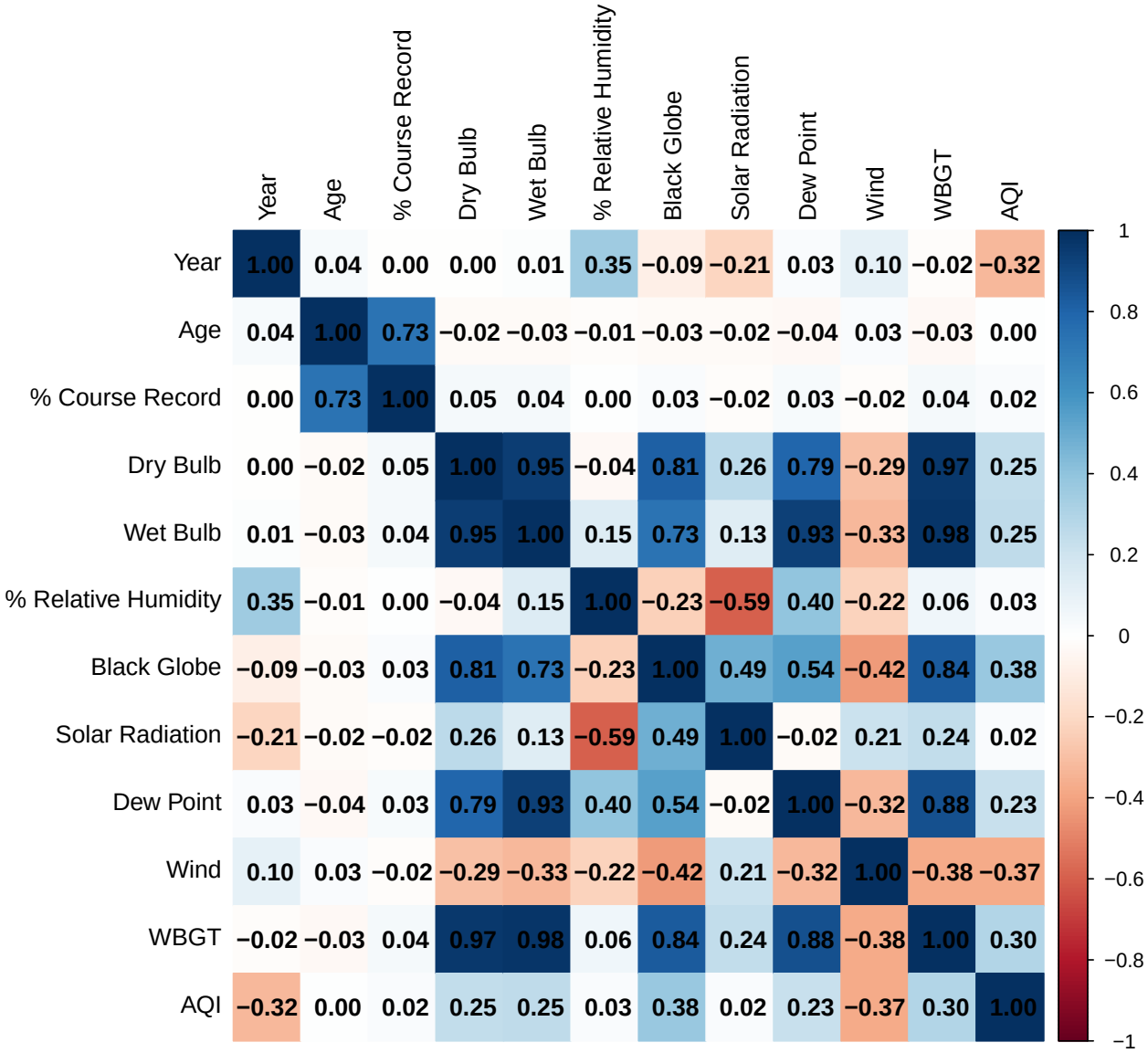

Figure 5: Correlations of Weather Variables with % off Course Record

We thought that stratifying the results by sex and age group might result in higher correlations of percent off of course record with the weather variables. In this correlation plot, we can see that the weather variables are not strongly correlated with percent off of course record for any of the groups. The performance of senior men is not correlated with any weather variables, but it is highly correlated with age (correlation = .80). Potentially, the effect of age is strong enough that it obscures the impact of any of the weather variables on percent off of course record. Senior women's performance is also not correlated with any weather variables, other than a weak (.12) correlation with AQI. Older men and women's performances are weakly correlated with dry bulb, wet bulb, black globe, dew point, WBGT, and AQI. Correlation does not seem to vary across gender within this age group. The runners in the middle age group's performances are also weakly correlated with the previously mentioned weather variables, except for AQI, although to a lesser extent than the "older" age group. These correlations also do not seem to vary across genders. We did not include the "Young" age group, since they had few observations and we suspected that they would not differ strongly from the "Middle" age group. Overall, WBGT seems to be the most correlated with a decreased performance across the age groups. We show this again in Figure 7, where the % off of course record increases as WBGT increases for the "young", "middle", and "older" age groups. We do not see a clear trend in the "senior" age group.

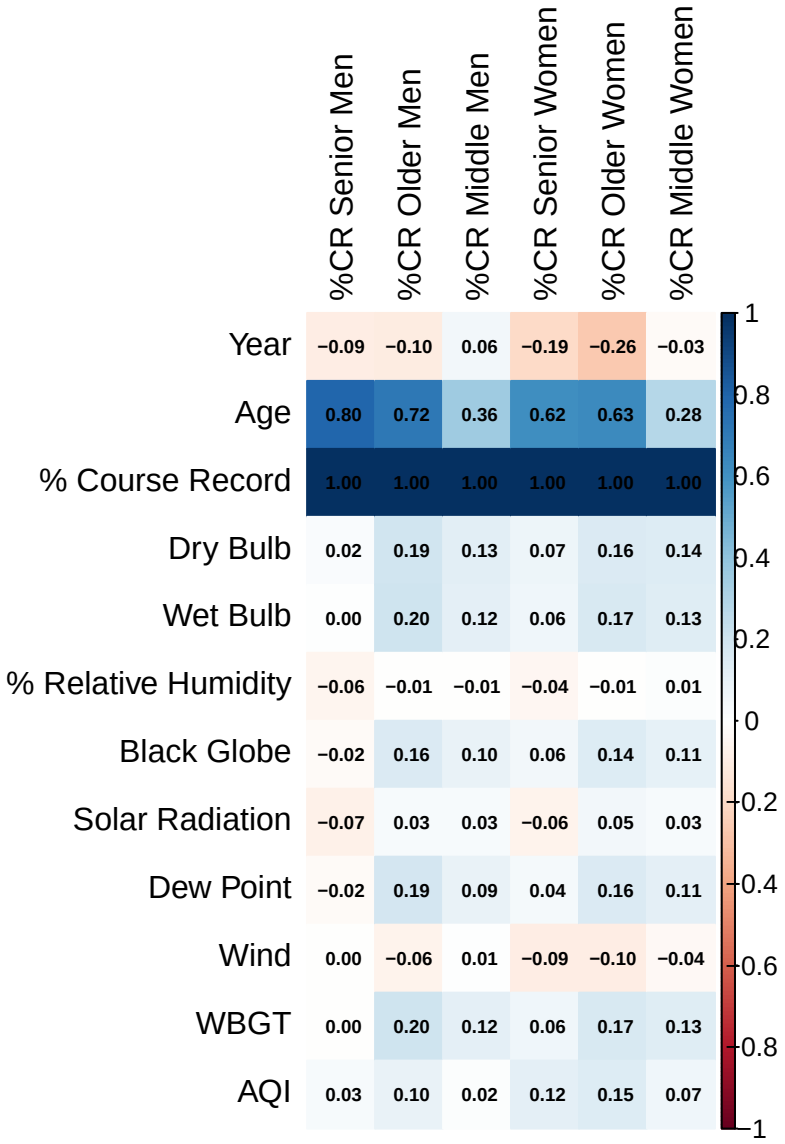

Figure 6: Correlations of Weather Variables with % off Course Record by Age Group and Gender

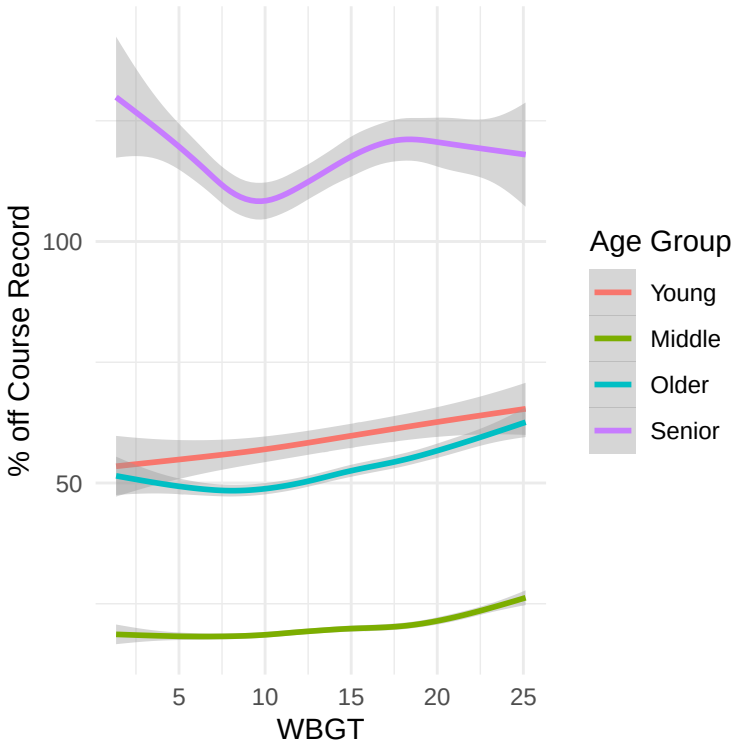

Figure 7: % off Course Record by WBGT by Age Group

Finally, to further investigate the effects of age, sex, and WBGT on marathon performance, we performed a linear regression. We first looked at the distribution of the variables in the model: percent off of course record, age, WBGT, and sex. We found that taking the log of the percent off of course record produced a more normal distribution, so we used logged percent off course record as our outcome. We had a three way interaction of age, sex, and WBGT as predictors, along with the two way interaction and main effects. The results of this regression are shown in Table 3. The interaction of age and WBGT is significant ($p < .05$), but the three way interaction is not, and neither is sex with WBGT. This agrees with the results from our previous correlation plots. The main effects of age, sex, and WBGT are significant, and so is the interaction of age with sex and age with WBGT.

We also performed a second linear regression (Table 4), this time using all the weather variables, age, year, race, and interactions between age with dry bulb temperature, relative humidity, solar radiation, and air quality index. We found that age, sex, race, relative humidity, solar radiation, air quality, the interaction of age with relative humidity, age with solar radiation, age with dry bulb, and age with air quality were significant.

Table 3: Results of Linear Regression of log(% Off of Course Record) ~ Interaction between Age:Sex:WBGT

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.251	0.077	16.193	0.000
Age	0.046	0.002	28.538	0.000
Sex	-0.446	0.105	-4.258	0.000
WBGT	0.045	0.005	8.289	0.000
Age:Sex	0.004	0.002	1.993	0.046
Age:WBGT	-0.001	0.000	-5.871	0.000
Sex:WBGT	0.003	0.007	0.359	0.719
Age:Sex:WBGT	0.000	0.000	0.118	0.906

Table 4: Results of Linear Regression of log(% Off of Course Record) ~ Weather Variables and Age

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.944	5.158	-0.377	0.706
Age	0.051	0.003	19.678	0.000
Year	0.001	0.003	0.538	0.591
RaceChicago	0.184	0.032	5.829	0.000
RaceGrandma's	0.086	0.050	1.721	0.085
RaceNew York	-0.009	0.036	-0.246	0.806
RaceTwin Cities	0.138	0.037	3.672	0.000
Sex	-0.200	0.018	-11.099	0.000
Dry Bulb	0.025	0.015	1.644	0.100
Wet Bulb	0.049	0.030	1.612	0.107
Relative Humidity	0.005	0.001	4.546	0.000
Black Globe	-0.007	0.003	-2.217	0.027
Solar Radiation	0.001	0.000	2.744	0.006
Dew Point	-0.023	0.013	-1.796	0.073
AQI	-0.006	0.001	-3.966	0.000
Age:Dry Bulb	-0.001	0.000	-5.629	0.000
Age:Relative Humidity	0.000	0.000	-4.751	0.000
Age:Solar Radiation	0.000	0.000	-2.321	0.020
Age:AQI	0.000	0.000	4.493	0.000

Discussion and Conclusion

This analysis shows that aging clearly impacts marathon performance, as shown in previous studies, and that this effect differs across gender. We initially hypothesized that aging would impact performance, but not necessarily across gender.

Furthermore, we show that the effects of weather conditions on marathon performance differ by age, when age is binned and when age is continuous. The results from Figure 4 demonstrate how the impact of the warning flag differs across age groups. Some weather conditions were shown to be important in predicting performance declines. These variables were relative humidity, solar radiation, and air quality. Since WBGT factors into these conditions, WBGT seems to be a good overall measure of weather conditions. However, since air quality is shown to be associated with marathon performance, it might be worthwhile to investigate how it could be factored into this the WBGT measurement, or how it should be included in further statistical analyses.

A limitation of this study is that the effects of the weather on marathon performance may not be entirely physiological, and that runners are aware of the weather conditions due to reading weather reports or seeing the flag conditions. Therefore, they might not be running at their fullest effort in order to protect their health and comfort. However, since this data contains only the top performers of each age, it is unlikely that the runners would be purposefully slowing themselves down. Another limitation is that the air quality index data is not as specific as it could be. Further research could map which exact air quality testing sites are nearest to the race course, rather than just using the closest CBSA code as we did in this study. Also, since this study is only looking at the top performers in each age group, it isn't necessarily applicable to slower or less experienced runners. Future research could focus on middle of the pack runners in each age group. Finally, since this is just an exploratory data analysis, more rigorous statistical analysis should be done to make any strong conclusions on the effects of weather on marathon runners.

In conclusion, aging is associated with decreased marathon performance in runners, and this effect varies across gender. Furthermore, weather conditions, specifically relative humidity, solar radiation, and aqi, affect marathon performance, and age interacts with all of these variables. The effects of weather on performance do not seem to differ by gender.

References

Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on Marathon-running performance. Medicine & Science in Sports & Exercise, 39(3), 487–493. https://doi.org/10.1249/mss.0b013e31802d3aba

Kenney, W. L., & Munce, T. A. (2003). Invited review: Aging and human temperature regulation. Journal of Applied Physiology, 95(6), 2598–2603. https://doi.org/10.1152/japplphysiol.00202.2003

Knechtle, B., McGrath, C., Goncerz, O., Villiger, E., Nikolaidis, P. T., Marcin, T., & Sousa, C. V. (2021). The role of environmental conditions on Master marathon running performance in 1,280,557 finishers the 'new york city marathon' from 1970 to 2019. Frontiers in Physiology, 12. https://doi.org/10.3389/fphys.2021.665761

Mccrowey C, Sharac T, Mangus N, Jager D, Brown R, Garver D, Wells B, Brittingham H (2023). _A R Interface to the US EPA Air Quality System Data Mart API_. <https://cran.r-project.org/package=RAQSAPI>.

Reusser, M., Sousa, C. V., Villiger, E., Alvero Cruz, J. R., Hill, L., Rosemann, T., Nikolaidis, P. T., & Knechtle, B. (2021). Increased Participation and Decreased Performance in Recreational Master Athletes in "Berlin Marathon" 1974-2019. Frontiers in physiology, 12, 631237. https://doi.org/10.3389/fphys.2021.631237

NOAA (National Oceanic and Atmospheric Administration). (2024). Heat stress datasets and documentation (provided to EPA by NOAA in April 2024) [Data set].

Code Appendix

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

#packages used for this rmd document, not including the RAQSAPI package used to
#retrieve data for air quality

library(knitr)
library(tidyverse)
library(lubridate)
library(corrplot)
library(gtsummary)
library(cowplot)
library(kableExtra)
#load in the data that we use throughout the report
data <- read_csv("/Users/rachelyost/Downloads/project1.csv")
course_record <- read_csv("/Users/rachelyost/Downloads/course_record (2).csv")
aqi_values <- read_csv("/Users/rachelyost/Downloads/aqi_values.csv")
#data preprocessing

#make column names easier to use within the code
colnames(data) <- c("Race", "Year", "Sex", "Flag", "Age", "CR", "TdC", "TwC",
                    "rh", "TgC", "SRWm2", "DP", "Wind", "WBGT")

#make sure columns are of the correct type
#check data types
str(data)

#make variables are factors if they should be factors
data$Race <- factor(data$Race)
data$Year <- factor(data$Year)
data$Sex <- as.factor(data$Sex)
data$Flag <- as.factor(data$Flag)

#Change names of races to actual names rather than numbers and change sex
#to Male/Female
data <- data %>% mutate(Race = case_when(Race == 0 ~ "Boston",
                                 Race == 1 ~ "Chicago",
                                 Race == 2 ~ "New York",
                                 Race == 3 ~ "Twin Cities",
                                 Race == 4 ~ "Grandma's"),
                        Sex = ifelse(Sex == "0", "Female", "Male"))

#Change course record data to match main data so it can be easily joined

course_record <- course_record %>%
  mutate(Race = case_when(Race=="B" ~ "Boston" ,
                          Race == "C" ~ "Chicago",
                          Race == "NY" ~ "New York",
                          Race == "TC" ~ "Twin Cities",
                          Race == "D" ~ "Grandma's"),
                          Gender = ifelse(Gender == "M", "Male", "Female")) %>%
```

```
  rename(Sex = Gender)

#make year a factor to join with main data set
course_record$Year <- factor(course_record$Year)

#join data with course record data and change course record to be in minutes
data <- data %>% full_join(course_record, data, by = c("Race", "Year", "Sex")) %>%
  mutate(CR.y = hms(CR.y))
data$CR.y = hour(data$CR.y)*60 + minute(data$CR.y) + second(data$CR.y)/60

#create a new variable Time, that shows the actual time for each participant
data$Time = ((data$CR.x/100)+1)*(data$CR.y)


#filter aqi_values to only include results with sample duration = 24 HOUR
aqi_dat <- aqi_values %>% filter(sample_duration == "24 HOUR")

#remove duplicates
include <- !duplicated(aqi_dat)
aqi_dat <- aqi_dat[include,]

#change format of year so that it matches the main data
aqi_dat$Year <- format(as.Date(aqi_dat$date_local, format="%Y/%m/%d"),"%Y")

#get mean aqi by cbsa_code and year and marathon
aqi_dat <- aqi_dat %>% group_by(cbsa_code, Year, marathon) %>%
  summarize(aqi= mean(aqi))

#remove cbsa code
aqi_dat <- aqi_dat[,-1]

#change column names for easy joining
colnames(aqi_dat) <- c("Year","Race","aqi")

aqi_dat <- aqi_dat %>% mutate(Race = case_when(Race == "Boston" ~ "Boston",
                                    Race == "Chicago" ~ "Chicago",
                                    Race == "NYC" ~ "New York",
                                    Race == "Twin Cities" ~"Twin Cities",
                                    Race == "Grandmas" ~ "Grandma's",
                                  ))

#merge the data
data_w_aqi <- left_join(data, aqi_dat)

#remove this empty entry
data_w_aqi <- data_w_aqi %>% filter(!(Race == "New York" & Year == "2012"))


#create bins for age groups
data_w_aqi <- data_w_aqi %>% mutate(Age_binned = case_when(Age < 20 ~ "Young",
                                        Age < 50 ~ "Middle",
                                        Age < 65 ~ "Older",
                                        Age < 99 ~ "Senior"))
```

```r
#order the flags and age binds correctly
data_w_aqi$Flag <- factor(data_w_aqi$Flag, levels =
                            c("White", "Green", "Yellow", "Red", "Black"))
data_w_aqi$Age_binned <- factor(data_w_aqi$Age_binned, levels =
                                  c("Young", "Middle", "Older", "Senior"))

#make Cr.x just CR

colnames(data_w_aqi)[6] <- "CR"
#examine missingness in the main data
sum(is.na(data_w_aqi))

#see which columns are missing data and how much
colSums(is.na(data_w_aqi))

#shows percent of data missing for each race and year to see if missingness
#is race-specific
summary_of_missing <- data_w_aqi %>%
  group_by(Race, Year) %>%
  summarise(across(everything(), ~ sum(is.na(.x))*100/n(), .names = "{.col}"))

summary_of_missing
#the data for the weather columns other than aqi is missing for race 1 in 2011,
#race 2 in 2011, race 3 in 2011, and race 4 in 2012
####get a histogram of the distribution of the values of each column
for (i in 1:ncol(data_w_aqi)){
  if (is.numeric(data_w_aqi[[i]])){
    hist(data_w_aqi[[i]], main = colnames(data_w_aqi)[i])
  }
  else{}
}

sum(na.omit(data$rh < 1))/nrow(na.omit(data))

min(na.omit(data[data$rh > 1,"rh"]))

#get humidity for each race
humidities <- data %>% group_by(Year, Race) %>% summarise(mean_humidity = mean(rh))

#if humidity was recorded as a decimal, change to a percent
data$rh <- ifelse(data$rh <= 1, data$rh*100, data$rh)

#remove variables that I don't want in table
data_for_summary_table <- data_w_aqi %>% dplyr::select(-"Year", -"CR.y")

#rename variables for the table
colnames(data_for_summary_table) <- c("Race", "Sex", "Flag", "Age","% Course Record",
                                      "Dry Bulb" ,"Wet Bulb", "% Relative Humidity"
                                      ,"Black Globe", "Solar Radiation" ,"Dew Point",
                                      "Wind",  "WBGT", "Time", "AQI", "Age Binned")

data_for_summary_table_races <- data_for_summary_table %>% select(Race, Flag, "Dry Bulb", "Wet Bulb",
```

```r
#don't include missing counts in table
tbl_summary(data_for_summary_table_races, by = Race, missing = "no") %>%
   as_kable_extra(booktabs = TRUE,
                  caption = "Race Characteristics",
                  longtable = TRUE, linesep = "") %>%
   kable_styling(font_size = 10,
                  latex_options = c("repeat_header", "HOLD_position"))

data_for_summary_table_people <- data_for_summary_table %>% select("Race", "Sex", "Age","% Course Rec

#don't include missing counts in table
tbl_summary(data_for_summary_table_people, by=Race, missing = "no") %>%
   as_kable_extra(booktabs = TRUE,
                  caption = "Participant Characteristics",
                  longtable = TRUE, linesep = "") %>%
   kable_styling(font_size = 10,
                  latex_options = c("repeat_header", "HOLD_position"))
min_times <- data_w_aqi %>% group_by(Race, Sex, Age) %>% summarise(meanCR =mean(CR)) %>% group_by(Rac

mean_times <- data_w_aqi %>% group_by(Race, Sex, Age) %>% summarise(meanCR =mean(CR))

times <- inner_join(min_times, mean_times) %>% filter(min_CR == meanCR)

#plot % off CR by age faceted by gender

ggplot(data_w_aqi, aes(x = Age, y = CR)) +
  scale_color_brewer(palette = "Set1") +
  geom_smooth(aes(color = Sex), se = TRUE, size = .5) +
  facet_wrap(~Race) +
  labs(x = "Age", y = "% off Course Record", color = "Sex", linetype = "Sex") +
  theme_minimal() +
  theme(
    text = element_text(family = "sans", color = "black"),
    strip.text = element_text(size = 14),
    axis.text = element_text(size = 7),
    axis.title = element_text(size = 14),
    legend.text = element_text(size = 14),
    legend.title = element_text(size = 14)
  ) +
  scale_x_continuous(breaks = seq(15, 90, 10))

##for race, find the lowest %CR

winners <- data_w_aqi %>% group_by(Race,Year,Sex) %>% filter(CR == min(CR))

#winners %>% group_by(Sex) %>% summarize(mean_age = mean(Age))
ggplot(winners) +
  geom_histogram(aes(x=Age, fill = Sex), color = "black", binwidth = 1) +
  labs(x = "Age", y = "Count") +
  facet_wrap(~Sex) +
  theme_minimal() +
  theme(text = element_text(family = "sans", color = "black"),
```

```
        strip.text = element_text(size = 14),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 14)) +
  scale_x_continuous(breaks = seq(min(winners$Age), max(winners$Age), by = 2)) +
  theme(legend.position="none") +
  scale_fill_brewer(palette = "Set1")

###look at how the fastest time for each age decreased by year
data_4_plot <- data_w_aqi %>%
  group_by(Sex,Year) %>%
  summarize(Time = mean(Time))
#plot marathon time in minutes by year
ggplot(data_4_plot, aes(x=as.numeric(as.character(Year)), y= Time)) +
  geom_point() + labs(x="Year", y = "Mean Race Time in Minutes") +
  geom_smooth(se = TRUE, color="black") +
  facet_wrap(~Sex) +
  theme_minimal() +
  scale_x_continuous(breaks =
                     seq(1993, 2016, by = 2)) +
  scale_y_continuous(breaks = seq(180, 235, by = 5)) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
###look at average times in each group for each flag

flag_age <- data_w_aqi %>% group_by(Age_binned, Flag, Race) %>%
  summarize(time = mean(Time), n = n(), se = sd(Time)/sqrt(n()))

flag_age <- na.omit(flag_age)

p_boston <- (ggplot(flag_age[flag_age$Race == "Boston",],
                    aes(x=Flag, y = time, fill=Flag)) +
               geom_col(color="Black")  +
               geom_errorbar(aes(ymin=(time - 1.96*se),
                                 ymax = (time + 1.96*se)), width = .5) +
              labs(y = "Mean Time", title = "Boston") +
              scale_fill_manual(values = c("White", "Green", "Yellow", "Red"))  +
              facet_wrap(~Age_binned)) +
              theme_minimal() +
              theme(legend.position="none",
                    axis.text.x = element_text(size = 8, angle = 30))

p_chicago <- (ggplot(flag_age[flag_age$Race == "Chicago",],
                     aes(x=Flag, y = time, fill=Flag)) +
              geom_col(color="black")  +
              geom_errorbar(aes(ymin=(time - 1.96*se),
                                ymax = (time + 1.96*se)), width = .5) +
              labs(y = "Mean Time", title = "Chicago") +
              scale_fill_manual(values = c("White", "Green", "Yellow", "Red")) +
              facet_wrap(~Age_binned)) +
              theme_minimal() +
              theme(legend.position="none",
                    axis.text.x = element_text(size = 8, angle = 30))
```

```
p_twin_cities <- (ggplot(flag_age[flag_age$Race == "Twin Cities",],
                         aes(x=Flag, y = time, fill=Flag)) +
                  geom_col(color="Black")  +
                  geom_errorbar(aes(ymin=(time - 1.96*se),
                                    ymax = (time + 1.96*se)),
                                width = .5) +
                  labs(y = "Mean Time", title = "Twin Cities") +
                  scale_fill_manual(values = c("White", "Green",
                                               "Yellow", "Red"))  +
                  facet_wrap(~Age_binned)) +
                  theme_minimal() +
                  theme(legend.position="none",
                        axis.text.x = element_text(size = 8, angle = 30))


plot_grid(p_boston, p_chicago, p_twin_cities)


#make year an integer for correlations
data_w_aqi$Year <- as.integer(as.character(data_w_aqi$Year))
#make sex numeric
data_w_aqi$Sex <- ifelse(data_w_aqi$Sex == "Female", 0, 1)

#### get correlations between %CR and each of the weather variables
#remove variables we aren't interest in
cor_data <- data_w_aqi %>% dplyr::select(-Race, -Flag, -Age_binned, -Time, -CR.y,
                                         -Sex)
cor_matrix <- cor(na.omit(cor_data), method = "spearman")
colnames(cor_matrix) <- c("Year" ,"Age" ,  "% Course Record" , "Dry Bulb",
                          "Wet Bulb"  , "% Relative Humidity" ,"Black Globe",
                          "Solar Radiation" ,"Dew Point" ,   "Wind",  "WBGT" ,"AQI")
rownames(cor_matrix) <- c("Year","Age" ,  "% Course Record" , "Dry Bulb",
                          "Wet Bulb"  , "% Relative Humidity" ,"Black Globe",
                          "Solar Radiation" ,"Dew Point" ,   "Wind",  "WBGT", "AQI")

#plot results with corrplot
corrplot(cor_matrix, method = 'color', addCoef.col = 'black', tl.col = "black",
         number.cex = 1, tl.cex = 1)


###compare senior men, senior women, with middle men, middle women
#make corr plots for each group

cor_data_senior_men <- data_w_aqi %>%
  dplyr::select(-Race, -Flag, -Time, -CR.y) %>%
  filter(Age_binned == "Senior", Sex== "1") %>%
  dplyr:: select(-Age_binned, -Sex)

cor_data_senior_women <- data_w_aqi %>%
  dplyr::select(-Race, -Flag, -Time, -CR.y) %>%
  filter(Age_binned == "Senior", Sex== "0") %>%
  dplyr:: select(-Age_binned, -Sex)
```

```r
cor_data_older_men <- data_w_aqi %>%
  dplyr::select(-Race, -Flag, -Time, -CR.y) %>%
  filter(Age_binned == "Older", Sex== "1") %>%
  dplyr:: select(-Age_binned, -Sex)

cor_data_older_women <- data_w_aqi %>%
  dplyr::select(-Race, -Flag, -Time, -CR.y) %>%
  filter(Age_binned == "Older", Sex== "0") %>%
  dplyr:: select(-Age_binned, -Sex)

cor_data_middle_men <- data_w_aqi %>%
  dplyr::select(-Race, -Flag, -Time, -CR.y) %>%
  filter(Age_binned == "Middle", Sex== "1") %>%
  dplyr:: select(-Age_binned, -Sex)

cor_data_middle_women <- data_w_aqi %>%
  dplyr::select(-Race, -Flag, -Time, -CR.y) %>%
  filter(Age_binned == "Middle", Sex== "0") %>%
  dplyr:: select(-Age_binned, -Sex)

cor_matrix_sm <- cor(na.omit(cor_data_senior_men))
cor_matrix_sw <- cor(na.omit(cor_data_senior_women))
cor_matrix_mm <- cor(na.omit(cor_data_middle_men))
cor_matrix_mw <- cor(na.omit(cor_data_middle_women))
cor_matrix_om <- cor(na.omit(cor_data_older_men))
cor_matrix_ow <- cor(na.omit(cor_data_older_women))

#create matrix
cor_matrix2 <- as.matrix(data.frame(cor_matrix_sm[,"CR"],cor_matrix_om[,"CR"],
                                    cor_matrix_mm[,"CR"],cor_matrix_sw[,"CR"],
                                    cor_matrix_ow[,"CR"],cor_matrix_mw[,"CR"]))

#change row and column names
colnames(cor_matrix2) <- c("%CR Senior Men", "%CR Older Men", "%CR Middle Men",
                           "%CR Senior Women", "%CR Older Women", "%CR Middle Women")


rownames(cor_matrix2) <- c("Year","Age" ,"% Course Record" , "Dry Bulb" ,
                           "Wet Bulb"  , "% Relative Humidity" ,  "Black Globe",
                           "Solar Radiation" ,"Dew Point" ,   "Wind","WBGT","AQI")

#plot results with corrplot
corrplot(cor_matrix2, method = 'color', addCoef.col = 'black', tl.col = "black",
         number.cex = .6, tl.cex = 1)

ggplot(data= data_w_aqi, aes(x=as.numeric(WBGT), y = as.numeric(CR), color = Age_binned)) + geom_smoo
#look at histograms of variables to check distributions
hist((data_w_aqi$Age))
hist(sqrt(data_w_aqi$Age))
hist(as.numeric(data_w_aqi$WBGT))
hist(as.numeric(data_w_aqi$CR))
hist(log(data_w_aqi$CR))
summary <- summary(lm(log(CR) ~ Age*Sex*WBGT, data=data_w_aqi))
```

```
kable(summary[["coefficients"]], caption = "Results of Linear Regression
      of log(% Off of Course Record) ~ Interaction between Age:Sex:WBGT", digits = 3)
summary2 <- summary(lm(log(CR) ~ Age + Year + Race + Sex + TdC + TwC + rh + TgC +
                         SRWm2 + DP + aqi + Age*TdC + Age*TdC + Age*rh +
                         Age*SRWm2 + Age*aqi, data=data_w_aqi))

summary2 <- summary2[["coefficients"]]

rownames(summary2) <- c("(Intercept)", "Age", "Year", "RaceChicago", "RaceGrandma's","RaceNew York","

kable(summary2, caption = "Results of Linear Regression
      of log(% Off of Course Record) ~ Weather Variables and Age", digits = 3)
```