PDA Project 1

Maia Lindner-Liaw

2024-10-01

Introduction

The marathon, a 42km (26.2 mile) footrace, is an endurance event requiring peak athletic performance from participants for an extended period of time[2]. In terms of performance by sex, women tend to experience less fatigue during endurance running and more even pacing, showing an advantage in ultra-running among women, although these effects are mitigated by other factors over shorter distances[4]. Given that these races occur outside, significant factors that could affect performance are weather conditions during the race. Studies have shown that performance in endurance events decreases as weather gets warmer, an effect which is magnified as distances and durations increase[1][2]. However, these studies only assessed elite male runners and did not quantify the amount of performance dropoff. Females and males have physiological differences that may affect thermoregulation, with some evidence that women tolerate hot and humid conditions better than men, but tolerate hot and dry conditions less well[5]. Age may also affect thermoregulation, as older people experience factors such as reduced sweat output and skin blood flow, which could make their thermoregulation less efficient[3]. Therefore, weather conditions may affect marathon performance in different amounts as age and sex varies, but this has not been rigorously studied. To start, we are interested in evaluating the effect of age on marathon performance for both men and women and identify peak competitive performance. Then we will look at how performance changes under differing weather conditions and whether the effects differ by sex and age. Variables we are interested in are temperature, humidity, solar radiation, wind, and dew point. Additionally, air quality has been shown to affect cardiorespiratory performance in older adults when exercising outside[6], so we will also investigate its effect on marathon performance. Finally, we are interested in determining which weather variables have the largest effect on marathon performance. Based on the existing literature, the hypothesis for this study is that older runners would be more affected by worse weather conditions than younger runners, and that the effects would be similar between men and women.

EDA

In this dataset, the best performances for each individual age are recorded for females and males at 5 US marathons spanning 1993-2016. The 5 races are the Chicago, New York City, Twin Cities, and Grandma's marathons. The Twin Cities and Grandma's marathons have data starting in 2000, Chicago has data starting in 1996, Boston has data starting in 1998, and New York City has data starting in 1993. All races have data up to the year 2016. The youngest runner across all races and years was 14 and the oldest was 91. Finishing times are converted to the percent off the course record to make the data comparable across races. For example, a percent course record of 40% means that person had a finishing time 1.4 times higher than the course record for that race. In addition the time measure of percent course record, the finishing times for each observation were converted to minutes using the available course record data for each race and year. This allows for a more natural finishing time that is easier to interpret. Variables such as the race variable, recorded as a numeric categorical, were modified to words to easier interpretation.

Weather data was collected for each race from Air Force sources. Weather variables included dry bulb temperature (TD), wet bulb temperature (TW), percent relative humidity (RH), black globe temperature (TG), solar radiation (SR), dew point (DP), and wind speed. Temperatures are measured in Celsius, with wet bulb accounting for humidity, and black globe accounting for solar radiation. Two summary variables are also included: wet bulb globe temperature (WBGT) and flag. WBGT is a weighted average of dry, wet, and black

Table 1: Runner Characteristics by Race

Characteristic	Boston N = 1886	Chicago N = 2191	Grandma's N = 944	New York City N = 2040	Twin Cities N = 1409
Age	47 (17)	46 (18)	44 (18)	50 (19)	45 (17)
Sex					
F	893 (47%)	1,041 (48%)	442 (47%)	981 (48%)	653 (46%)
M	993 (53%)	1,150 (52%)	502 (53%)	1,059 (52%)	756 (54%)
Percent_CR	42 (34)	51 (46)	46 (38)	54 (55)	45 (36)
Time	189 (47)	200 (62)	201 (54)	208 (75)	199 (51)

[1] Mean (SD); n (%)

globe temperatures, the flag variable bins the WBGT into levels ranging from `White` indicating the coolest conditions (WBGT<10C), to `Black` where conditions are bad enough that races are canceled (WBGT>28C).

Additionally, air-quality index (AQI) data was scraped for the days and locations corresponding to the marathons in the data set. The three measures in the data set are "PM2.5-Local Conditions", "Acceptable PM2.5 AQI & Speciation Mass", and "Ozone", corresponding with the parameter codes 88101, 88502, and 44201, respectively[6]. PM2.5 is the measure of particulate matter 2.5 micrometers and smaller in the air, classified as fine inhalable particles[7]. The 88101 code, representing local PM2.5 conditions, has measurements made using specific devices that meet certain standards, allowing this measure to be used for regulatory purposes. The other PM2.5 measure, code 88502, uses similar devices that are not rated the same and may be less accurate, hence the "acceptable" portion of the name [8]. While both are used for general reporting, 88101 is more stringent due to its use in regulation, which is why why choose to use this PM2.5 measure in our analysis instead of 88502. The ozone measures are unrelated to PM2.5, so we will include that variable regardless. The data set contains many observed values of each variable for the same race, so we will take the daily average to get a general overview of the AQI. Because we do not have detailed information on times during the day that people are running, it would not be feasible to get more granular. This data is joined to the marathon data.

An important step before further analysis is to investigate missing values in the data. When looking at the proportion of missingness in each column, we find that only the weather variables have missingness, and they each have the same 4.25% missingness. This suggests that some races may be missing weather data, which is confirmed when seeing that the missingness is exclusive to the Chicago, NYC, and Twin Cities marathons in 2011, and Grandma's marathon in 2012. We have a large data set of 11564 observations with less than 5% missingness, and the missingness is restricted to races with fully missing weather data, so analysis is unlikely to be significantly biased. Therefore, complete case analysis is appropriate, and those races will be removed for weather analysis.

In checking the distributions of the variables, an irregularity in the relative humidity variable was noted. While most of the observations were between 30 and 100, representing a percentage, some were less than one. This is likely a data quality issue where relative humidity was recorded as a percentage for some observations, but a decimal for others. Therefore, we will convert the decimal observations to percentages.

Table 1 summarizes the characteristics of the runners for each race, and Table 2 summarizes the weather conditions for the different races.

Chicago and New York have more observations than the other marathons, but recall these are the races with data from 1996 and 1993, respectively. We can see that there are slightly more male observations than female races, with the proportion remaining similar across races. Average ages and standard deviations are similar across races, with NYC having the oldest average of 50 years. Boston is the fastest race, with the average finishing time 12-21 minutes faster than than the others, which is also reflected in the lowest average percent course record of 41%. This is unsurprising, as the runners in the Boston marathon must qualify by time to enter, biasing the race toward faster runners.

Table 2: Weather Characteristics by Race

Characteristic	**Boston** N = 1886	**Chicago** N = 2191	**Grandma's** N = 944	**New York City** N = 2040	**Twin Cities** N = 1409
Flag					
White	928 (49%)	608 (28%)	0 (0%)	1,146 (56%)	471 (33%)
Green	720 (38%)	1,347 (61%)	464 (49%)	770 (38%)	716 (51%)
Yellow	115 (6.1%)	120 (5.5%)	358 (38%)	124 (6.1%)	222 (16%)
Red	123 (6.5%)	116 (5.3%)	122 (13%)	0 (0%)	0 (0%)
Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
TD	11.7 (6.2)	12.8 (6.2)	17.8 (3.0)	11.0 (4.1)	12.1 (4.6)
TW	7.4 (4.0)	8.9 (5.8)	14.9 (1.9)	6.7 (3.9)	8.5 (4.6)
RH	58 (20)	61 (10)	75 (13)	53 (14)	59 (15)
TG	25 (8)	25 (7)	29 (9)	21 (6)	25 (6)
SR	635 (191)	462 (99)	636 (201)	394 (139)	452 (129)
DP	2.9 (4.4)	5.1 (6.9)	13.2 (2.1)	1.7 (5.2)	4.0 (6.5)
Wind	12.5 (4.4)	8.4 (3.3)	8.0 (3.0)	10.8 (5.0)	8.0 (2.5)
WBGT	11.4 (4.7)	12.5 (5.9)	18.1 (3.1)	10.0 (4.1)	12.1 (4.6)
Time	189 (47)	200 (62)	201 (54)	208 (75)	199 (51)
PM2.5	43 (16)	53 (18)	32 (12)	43 (21)	31 (17)
Ozone	47 (16)	33 (11)	31 (7)	27 (4)	26 (9)

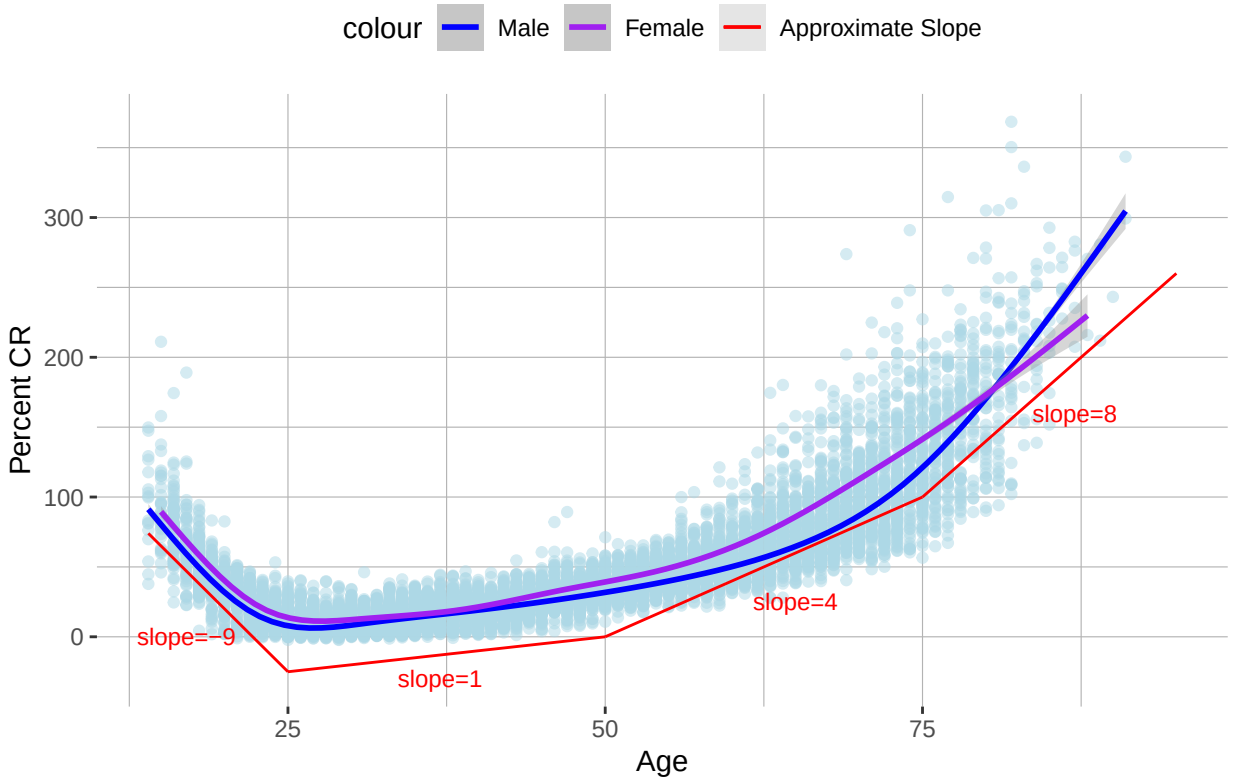
[1] n (%); Mean (SD)

Excluding wind, the weather measurements for the Boston, Chicago, NYC, and Twin Cities races are similar with moderate values and primarily white and green flags. In contrast, Grandma's race has 50% yellow flag rate with no white flag observations and an average WBGT measure 5.4C higher than the next highest average. This is not unexpected when considering that Grandma's is the only race run in the summer. Boston runs in April, Chicago and Twin Cities run in October, NYC runs in November, but Grandma's runs in the middle of June[9]. We would expect more temperate conditions in the spring and fall compared to the summer. PM2.5 is high in Chicago compared to the other locations, however, the average ozone is more moderate.

Aim 1

The first aim of this project is to assess the effect of age on performance for both men and women. Figure 1 plots percent course record against age with smoothed lines, separating by sex. In this aim, we are interested at looking at the differences in effect age has on both genders and not the raw differences between genders. Therefore, it makes more sense to use percent course record instead of time in minutes, as each observation will be compared to the best ever performance for that gender.

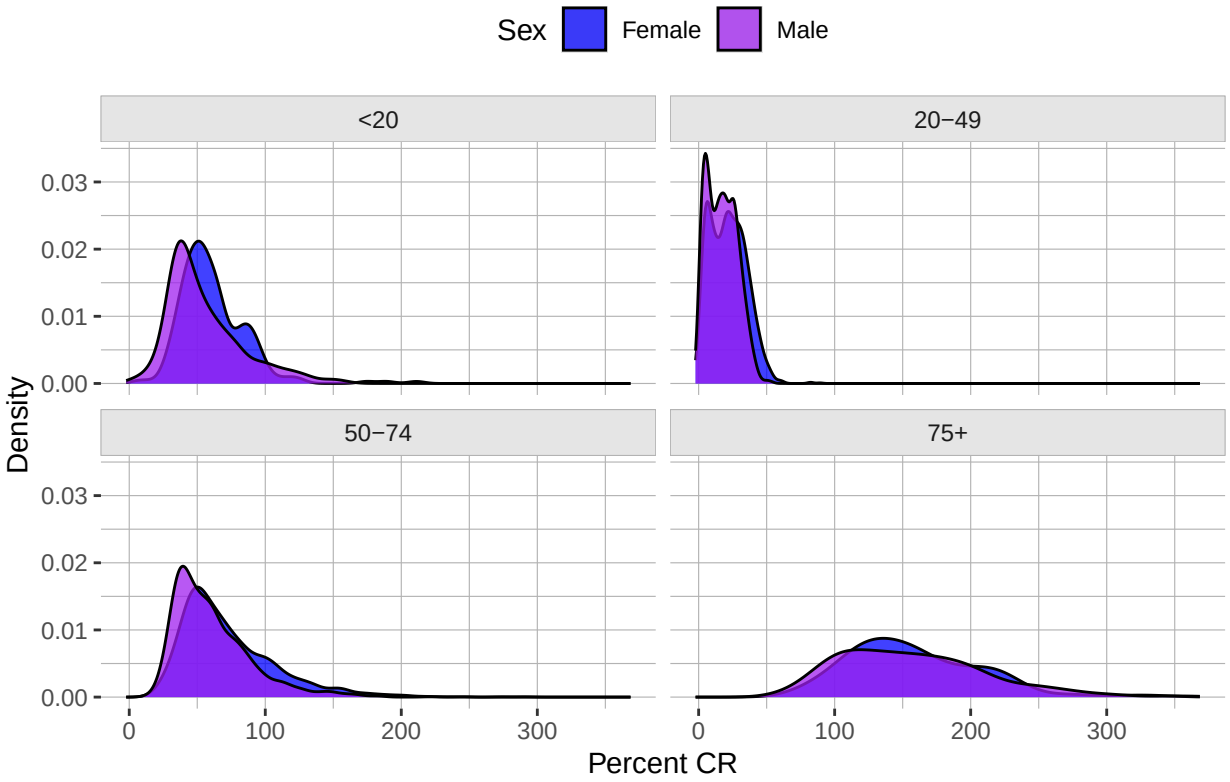
Figure 1: Finishing Times as Percent Course Record by Sex and Race



We can see that the data forms a U-shape, indicating a non-linear relationship between marathon performance and age. Peak performance for both sexes occurs just after age 25, as indicated by the lowest points on the smoothed lines. The smoothed lines are relatively flat between ages 25 and 50, indicating age does not have much effect in this age range, whereas the curves slope more for runners under 25 and above 50. Runners under 25 improve dramatically for each year older, and runners above 50 slow down increasingly quickly for each additionally year older. Overall, men have a slight performance advantage in terms of percent course record over women at most ages as expected, however at ages greater than approximately 80, women actually outperform men on average.

Based on the shape of the curve in Figure 1, we can select natural bins for age by performance. Grouping age to under 20, 25-49, 50-74, and 75+ will allow us to better look at the performance distributions at different ages. Figure 2 shows these distributions for both sexes.

Figure 2: Distribution of Finishing Times by Sex and Age Category



For all of the age groups, there is considerable overlap between the distributions for males and females, with the male distributions shifted slightly faster. The <20 age group has the least overlap, meaning that at ages under 20, males have more of a performance advantage over women compared to the other age groups.

We see that the distributions for the under 20 and 50-74 age groups are quite similar, peaking near 50% course record and almost all of the density falling between 0% and 150% course record. They both have long tails out to the right, indicating that there were some very slow runners, but not many of them. The 20-48 age group clearly has the best performance, with almost the entire density falling under 50% course record. This combined with the few very slow runners results in the extreme skew of the distribution, which is more than in any other age group.

The over 75 age group has a much flatter distribution than the other age groups due to the broader range of finishing times. The variance in finishing times is the smallest in the 20-49 group, largest in the 75+ age group, and somewhere in between for the <20 and 50-74 age groups.

Based on Figures 1 and 2, we can see age does affect marathon performance, but the effect is different based on age group. For runners under 25, increasing age improves marathon performance, whereas after 25, increasing age reduces performance. The amount of performance reduction is small between aged 25 and 50, but becomes increasingly larger for ages 50+. In Figure 1, we quantify this by approximating the slopes for four sections of the smoothed curves. For runners under 20, every unit increase in age results an approximate 9 percent improvement in percent course record on average. From ages 25 to 50, every unit increase in age results in an about equal increase in percent course record. From 50 to 75, being one year older results in an approximate slowing of 4 percent of the course record. Finally, aging 1 year for those over 75 leads to a slowing of about 8 percent of the course record. The peak performance years for both sexes is around 25-30 years old, and the corresponding 20-49 age group has the lowest variation in finishing times when excluding the few very slow runners. The 75+ age group has the slowest finishing times as we would expect, but also the highest variability in performance. In general, the effect of age does not greatly differ between males and females.

Table 3: Data Characteristics by Flag

Characteristic	**White** N = 3153	**Green** N = 4017	**Yellow** N = 939	**Red** N = 361	**p-value**
Race					<0.001
Boston	928 (29%)	720 (18%)	115 (12%)	123 (34%)	
Chicago	608 (19%)	1,347 (34%)	120 (13%)	116 (32%)	
Grandma's	0 (0%)	464 (12%)	358 (38%)	122 (34%)	
New York City	1,146 (36%)	770 (19%)	124 (13%)	0 (0%)	
Twin Cities	471 (15%)	716 (18%)	222 (24%)	0 (0%)	
Sex					>0.9
F	1,495 (47%)	1,900 (47%)	443 (47%)	172 (48%)	
M	1,658 (53%)	2,117 (53%)	496 (53%)	189 (52%)	
Age	47 (18)	47 (18)	45 (18)	46 (18)	0.045
Percent_CR	47 (46)	49 (44)	48 (41)	54 (41)	<0.001
Time	198 (62)	200 (60)	201 (56)	206 (56)	<0.001

[1] n (%); Mean (SD)

[2] Pearson's Chi-squared test; Kruskal-Wallis rank sum test

Aim 2

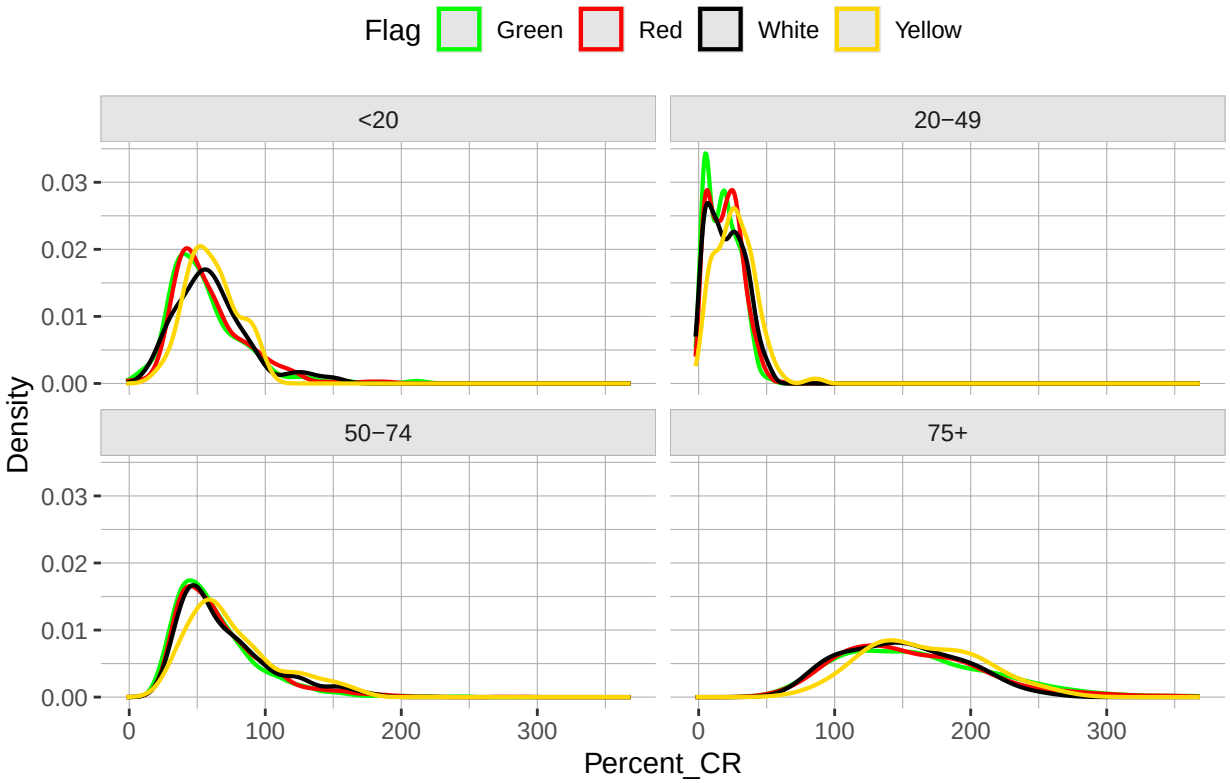
The second aim of this project is to investigate if and how weather conditions affect marathon performance, and if these effects differ by sex and age. Recall one of the summary weather variables was the flag status of each race. We can look at a summary of the data grouped by these flags to see if there are any differences between the groups. Table 3 shows these characteristics, excluding the raw weather measurements, as they are used to calculate flag status.

As we previously noted, Grandma's marathon is the only one running in the summer, and it is therefore unsurprising to see that it accounts for 38% and 34% of the yellow and red flags, respectively. The number of observations under a red flag is much smaller than all other flags, which is to be expected, as adverse weather conditions leads to fewer participants and finishers. There is no difference in gender of finishers across the different flags, with the proportions of male and female observations remaining almost exactly the same.

Age and finishing time, as both percent course record and time in minutes, both show significant differences across the flag types. The average age of finishers is lower under red and yellow flags than under white and green flags, and the slowest average finishing times occur under red flag conditions. Average finishing times under white and green flag conditions are the fastest, echoing the results of average ages under the best conditions. Based on this table, weather conditions do significantly affect marathon performance.

In Figure 3, we look at the how weather conditions affect age groups differently by using the same age categories as before. Figure 3 shows the distributions of percent course record for each flag type, separated into age bins.

Figure 3: Distribution of Finishing Times by Age and Flag

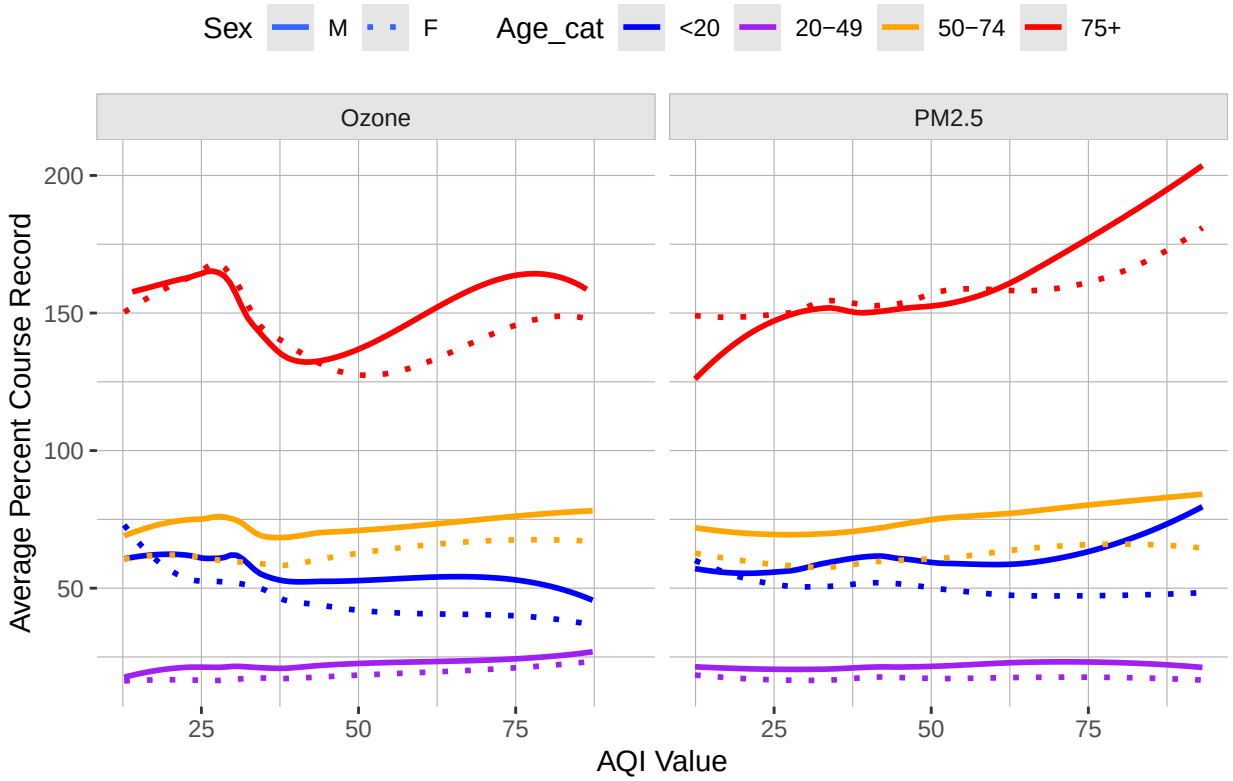


There are some visible differences in Figure 3, but they are slight. In the under 20 age group, we can see a clear difference in peak location for white and green flags compared to red and yellow flags. The peaks of the red and yellow distributions are clearly to the right of the white and green flag distributions, indicating slower times under the hotter conditions. In the 50-74 age group, the distribution of finishing times under yellow flag conditions is more similar to white and green flag conditions, and the distribution for red flag conditions is slightly off to the right. Similarly in the 20-49 age group, there is not much difference between the yellow flag distribution compared to the white and green flag conditions. Again, the red flag conditions distribution is slower than the other flags, but there is still considerable overlap. Finally, the 75+ age group distributions have high variance, resulting in flatter peaks, which makes comparison less clear. However, red and yellow flag condition distributions do appear further right, indicating slower finishing times.

Overall, there is not much visual difference between the distributions of finishing times for the different flag conditions. In all age groups, the red flag conditions resulted in slower finishing time. Under yellow flag conditions, only the under 20 and over 75 age groups show finishing time distributions further right, suggesting that the 20-49 and 50-74 age groups are not as affected by yellow flag conditions.

We will also look at the effect of AQI measurements which are separate measurements than those used to calculate WBGT.

Figure 4: Effect of AQI Measures on Finishing Times



In Figure 4, we see the average percent course record finishing times as AQI values increase, indicating worsening air quality. For the youngest three age groups, there does not appear to be much effect for either males or females. However, we do see some effect on the finishing times in the 75+ age group. As PM2.5 increases, the oldest group sees a slowing of finishing times, particularly once PM2.5 surpasses 60, with a slightly greater effect in females. For ozone levels, the relationship is unexpected. From ozone levels of about 30-40, finishing times actually improve, before slowing occurs again.

In Figures 3, we saw some differences in finishing times as weather conditions worsened, but the differences were small. Furthermore, conditions did not seem to affect genders differently, but there was some difference in effect across age groups. In Figure 4, we also see an increasing trend in average percent course record as conditions worsen, although the difference was mainly for the oldest group. These results agree with the findings in Table 3, but we can use one more measure to determine the extent of these differences. We will fit a linear regression model to finishing time with interaction terms between WBGT and age, and AQI and age. To better compare the effects by sex, we will fit separate models for males and females and compare the results. In this case, we will use finishing time in minutes, because that has a more natural interpretation than percent course record in this setting. Because we recognize that the 20-49 age group is the highest performing, we will set it as the baseline level. The results of these simple linear regressions are found in Tables 4 and 5.

In Tables 4 and 5, we see that age and weather conditions as summarized by WBGT all have significant effects on finishing time in minutes, with p-values all less than 0.001. Using the under 20-49 age group as the baseline, being under 20 results in a 65 minute slower in finishing time on average for men and 62 for women. Being 50-74 results in a 50 minute slower finishing time for men on average and 70 for women. Being being 75+ results in a 197 minute drop off in finishing time for women on average and 192 for men. For every unit increase in WBGT, finishing time increases by 0.66 minutes for women on average for the baseline group and 0.41 for men in the baseline group. Although these are different, the magnitude is small, which supports our previous findings that the effect doesn't differ much by sex. These interpretations hold all other variables constant. For both sexes, the AQI measures have no significant effect on finishing times.

Table 4: Linear Regression of Time (m) with WBGT Interactions in Females

Characteristic	**Beta**	**95% CI**	**p-value**
Age_cat			
20-49	—	—	
<20	62	45, 79	<0.001
50-74	70	62, 78	<0.001
75+	197	175, 219	<0.001
WBGT	0.66	0.33, 0.99	<0.001
Ozone	-0.08	-0.22, 0.05	0.2
PM2.5	-0.04	-0.13, 0.05	0.4
Age_cat * WBGT			
<20 * WBGT	0.41	-0.59, 1.4	0.4
50-74 * WBGT	0.38	-0.13, 0.89	0.14
75+ * WBGT	-0.85	-2.4, 0.64	0.3
Age_cat * Ozone			
<20 * Ozone	-0.51	-0.99, -0.03	0.039
50-74 * Ozone	-0.36	-0.56, -0.15	<0.001
75+ * Ozone	-1.0	-1.6, -0.45	<0.001
Age_cat * PM2.5			
<20 * PM2.5	0.15	-0.14, 0.45	0.3
50-74 * PM2.5	0.25	0.12, 0.39	<0.001
75+ * PM2.5	0.93	0.56, 1.3	<0.001

[1] CI = Confidence Interval

When looking at the interaction terms, we see some of the results are significant. The interaction terms for the categorical age and WBGT interaction are significant for men but not for women. Women of all ages are affected similarly by WBGT, but men are affected differently. WBGT has a significantly higher effect in under 20 and 50-74 age groups compared to the baseline group. For ozone and PM2.5, both sexes have significant interactions with age, even though the main effects are not significant. Overall, ozone had a negative effect on finishing times, with the effect being stronger in the under 20s and the over 75s for both sexes. The effect of PM2.5 increased as age increased, with older people experiencing a greater slowing effect.

Based on the tables and figures, we find that weather conditions do have a negative impact on marathon performance in general. The effect does not vary significantly between genders, but does between age. Compared to the baseline 20-49 age group, WBGT results in more slowing for age groups <20 and 50-74, but less slowing for ages 75+. Increasing ozone resulted in improvements in finishing times, with more prominent effects in the <20 and 75+ age groups. Increasing PM2.5 lead to slower times, with larger effects for older runners.

Aim 3

The final aim of this project is to determine which weather conditions have the largest impact on race performance. To determine this, we will again fit a simple linear regression model to assess the significance and magnitude of associations between weather conditions and finishing time. As before, we will use finishing time in minutes for better interpretation. Because WBGT is calculated from dry bulb, wet bulb, and black globe temperature, we will exclude those underlying measures and only include WBGT to prevent multicollinearity. This leaves the RH, SR, DP Wind, WBGT, Ozone, and PM2.5 variables to be included in the model. We further check multicollinearity by looking at the correlations between these variables. Figure 5 shows a large correlation between WBGT and DP, so we will not include DP in the model. We also can check for non-linearity in the variables by plotting them against finishing Time. When doing so, we see no

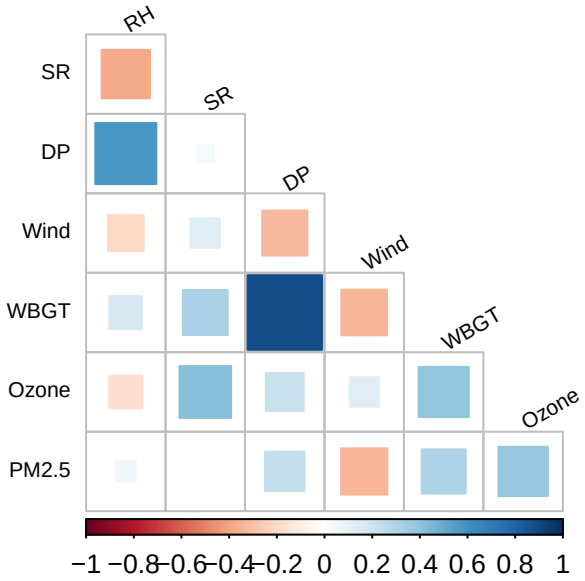
Table 5: Linear Regression of Time (m) with WBGT Interactions in Males

Characteristic	Beta	95% CI	p-value
Age_cat			
20-49	—	—	
<20	65	51, 80	<0.001
50-74	50	43, 57	<0.001
75+	192	180, 203	<0.001
WBGT	0.41	0.11, 0.70	0.007
Ozone	0.00	-0.12, 0.12	>0.9
PM2.5	-0.05	-0.13, 0.03	0.2
Age_cat * WBGT			
<20 * WBGT	0.77	-0.09, 1.6	0.078
50-74 * WBGT	0.59	0.15, 1.0	0.009
75+ * WBGT	-0.32	-1.1, 0.52	0.5
Age_cat * Ozone			
<20 * Ozone	-0.66	-1.1, -0.24	0.002
50-74 * Ozone	-0.20	-0.38, -0.02	0.033
75+ * Ozone	-1.1	-1.5, -0.83	<0.001
Age_cat * PM2.5			
<20 * PM2.5	-0.14	-0.39, 0.11	0.3
50-74 * PM2.5	0.11	-0.01, 0.23	0.067
75+ * PM2.5	0.67	0.45, 0.90	<0.001

[1] CI = Confidence Interval

obvious non-linearity in the weather variables, but recall the non-linear shape of Figure 1. Therefore, we will include continuous age with a squared term in the model but leave the rest linear. We will again fit separate models for sex to better assess the differences.

Figure 5. Correlation Between Weather Variables



Based on the results in Tables 6 and 7, we see that the most significant weather variables are relative

Table 6: Linear Regression of Time (m) and Weather Conditions in Females

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	309	289, 329	<0.001
RH	-0.07	-0.26, 0.11	0.4
SR	0.01	0.00, 0.01	0.036
DP	0.14	-0.89, 1.2	0.8
Wind	-0.05	-0.30, 0.20	0.7
WBGT	0.82	-0.30, 1.9	0.2
Ozone	-0.35	-0.43, -0.26	<0.001
PM2.5	0.10	0.05, 0.16	<0.001
Age	-8.1	-8.4, -7.8	<0.001
I(Age^2)	0.11	0.11, 0.12	<0.001

[1] CI = Confidence Interval

Table 7: Linear Regression of Time (m) and Weather Conditions in Males

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	293	276, 310	<0.001
RH	-0.15	-0.32, 0.01	0.069
SR	0.00	-0.01, 0.00	0.7
DP	0.61	-0.30, 1.5	0.2
Wind	-0.04	-0.26, 0.19	0.7
WBGT	0.14	-0.83, 1.1	0.8
Ozone	-0.21	-0.29, -0.14	<0.001
PM2.5	0.04	-0.01, 0.09	0.089
Age	-7.6	-7.8, -7.4	<0.001
I(Age^2)	0.10	0.10, 0.10	<0.001

[1] CI = Confidence Interval

humidity (RH), solar radiation (SR), Ozone, and PM2.5. Ozone is significant for both men and women, PM2.5 is significant for women and borderline significant for men, SR is significant for women only, and RH is borderline significant for men only. Increases in SR lead to slower times, increases in ozone lead to faster times, increases in PM2.5 lead to slower times, and increases in RH lead to faster times. The effects of ozone and RH are surprising in that they seem to improve finishing times, but the other effects are as we might expect. In Figure 4 we also saw that AQI had some effect on finishing time, so this model supports that conclusion. Table 4 showed that WBGT had a significant effect on performance even though the model in Tables 6 and 7 did not. Overall, weather factors WBGT, AQI, RH, and SR had the most significant effects on marathon performance.

Limitations

Overall, some of the results were expected, but some were also surprising. For example, we would not expect ozone and RH to improve performance. However, it is important to note that we used simple linear regression models and did not do thorough model selection or diagnostics. Although we did not detect obvious non-linearity in the variables when plotting them against finishing time, there may still be misspecification in the model. This means that our unexpected results may be the result of ill-fitting models, and the true effects are closer to what we would expect.

Another limitation of our data is that it contains only the top finisher for every age in each race. This may cause selection bias towards high performers. As we saw in Figure 4, weather conditions may not affect the elite athletes as much as the average athlete, so selecting only the best performers from each age may underestimate the effect of weather conditions. Because there are fewer observations for the very young and very old ages, we may not have adequate sample size to assess associations.

In terms of the weather data, the measurements are averages across the entire day, which may not accurately reflect the weather conditions of each runner. For example, if a runner started later, it may have gotten hotter, or if a slow runner spends longer on course, they may have experienced a longer period of hotter conditions. The weather data is not granular to the individual level, which also may affect the associations between weather conditions.

References

1. Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. Med Sci Sports Exerc, 42(1), 135-41.
2. Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. Medicine and science in sports and exercise, 39(3), 487-493.
3. Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.
4. Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., ... & Millet, G. Y. (2022). Sex differences in endurance running. Sports medicine, 52(6), 1235-1257.
5. Yanovich, R., Ketko, I., & Charkoudian, N. (2020).
6. Stieb DM, Shutt R, Kauri LM, Roth G, Szyszkowicz M, Dobbin NA, Chen L, Rigden M, Van Ryswyk K, Kulka R, Jovic B, Mulholland M, Green MS, Liu L, Pelletier G, Weichenthal SA, Dales RE. Cardiorespiratory Effects of Air Pollution in a Panel Study of Winter Outdoor Physical Activity in Older Adults. J Occup Environ Med. 2018 Aug;60(8):673-682. doi: 10.1097/JOM.0000000000001334. PMID: 29668530.
7. Environmental Protection Agency. (2024, February 13). AQS Code List. EPA. https://www.epa.gov/aqs/aqs-code-list
8. Environmental Protection Agency. (2024b, August 16). Particulate Matter (PM2.5) Trends. EPA. https://www.epa.gov/air-trends/particulate-matter-pm25-trends
9. Environmental Protection Agency. (2024a, January 4). What is the difference between parameter codes 88101 and 88502 for PM2.5 monitors?. EPA. https://www.epa.gov/outdoor-air-quality-data/what-difference-between-parameter-codes-88101-and-88502-pm25-monitors#:~:text=The%2088101%20monitors%20include%20both,not%20used%20for%20regulatory%20purposes.

10. USA Marathon List. US Marathon Calendar 2024-2025 | Marathons in America. (2024). http://www.usamarathonlist.com/

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
library(knitr)
library(tidyverse)
library(lubridate)
library(gtsummary)
library(kableExtra)
library(gt)
library(gridExtra)
library(corrplot)
# read course record data
cr_dat<-read.csv("/Users/mlindnerliaw/Desktop/course_record.csv")
# expand race names
cr_dat$Race<-case_when(cr_dat$Race=="B"~"Boston",
                       cr_dat$Race=="C"~"Chicago",
                       cr_dat$Race=="D"~"Grandma's",
                       cr_dat$Race=="NY"~"New York City",
                       TRUE~"Twin Cities")

# read data
dat<-read.csv("/Users/mlindnerliaw/Desktop/project1.csv")
# rename columns
names(dat)<-c("Race", "Year", "Sex", "Flag", "Age", "Percent_CR", "TD", "TW",
              "RH", "TG", "SR", "DP", "Wind", "WBGT")
dat$Sex<-ifelse(dat$Sex==1, "M", "F")
# label races by location
dat$Race<-case_when(dat$Race==0~"Boston",
                    dat$Race==1~"Chicago",
                    dat$Race==2~"New York City",
                    dat$Race==3~"Twin Cities",
                    TRUE~"Grandma's")



# join course record data by race, year, and gender
full_dat<-left_join(dat, cr_dat, by=c("Race", "Year", "Sex"="Gender"))

# convert CR to minutes
in_minutes<-function(times){
  #' takes a vector of times in the format "hours:minutes:seconds" and returns the number of minutes
  #' @param times a character vector containing the times to convert
  #' @return a numeric vector of the same length as times containing the times in minutes
  #'

  # convert to hours minutes and seconds
  convert<-hms(times)

  # calculate minutes
  mins<-hour(convert)*60+minute(convert)+second(convert)/60
  return(mins)
}
```

```r
#convert percent_cr to times
full_dat<-full_dat%>%mutate(cr.time=in_minutes(CR),
                            Time=cr.time*(1+Percent_CR/100))



# fix irregularity of decimal relative humidities by converting to percent
full_dat<-full_dat%>%mutate(RH=ifelse(RH<1, RH*100, RH))

# blank flag values are black missing
full_dat$Flag<-ifelse(full_dat$Flag=="", NA, full_dat$Flag)
full_dat$Flag<-factor(full_dat$Flag, levels = c("White", "Green", "Yellow", "Red","Black"))


# read aqi data
aqi<-read.csv("/Users/mlindnerliaw/Desktop/aqi_values.csv")
aqi<-aqi%>%group_by(date_local, marathon, parameter_code)%>%summarize(aqi=mean(aqi, na.rm=TRUE))%>%
  filter(parameter_code!=88502)%>%mutate(Year=year(date_local))
aqi$Race<-case_when(aqi$marathon=="Boston"~"Boston",
                    aqi$marathon=="Chicago"~"Chicago",
                    aqi$marathon=="NYC"~"New York City",
                    aqi$marathon=="Twin Cities"~"Twin Cities",
                    TRUE~"Grandma's")
pm2.5<-aqi[aqi$parameter_code==88101,c("Year", "Race", "aqi")]
colnames(pm2.5)[3]<-"PM2.5"
ozone<-aqi[aqi$parameter_code==44201,c("Year", "Race", "aqi")]
colnames(ozone)[3]<-"Ozone"

# join aqi data
full_dat<-full_dat%>%left_join(pm2.5, by=c("Year", "Race"))%>%left_join(ozone, by=c("Year", "Race"))

## numbers for summary in text
get_years<-function(data, race){
  #' this function finds the range of years for a given race in the data
  #' @param data, the dataframe containing at least the race and year columns
  #' @param race, character name of race
  #' @return vector containing the start and endpoints of the range of years for the given race
  #'
  data%>%filter(Race==race)%>%select(Year)%>%unique()%>%range()
}

# boston race years
get_years(full_dat, "Boston")
# NYC race years
get_years(full_dat, "New York City")
# TC race years
get_years(full_dat, "Twin Cities")
# Gma's race years
get_years(full_dat, "Grandma's")
# Chi race years
get_years(full_dat, "Chicago")
num_missing<-function(data){
  #' Returns the number of missing values in the given data
```

```
  #' @param data, data vector or data frame of any type
  #' @return the number of NA values in the data

  return(length(which(is.na(data))))
}

# table of missingness proportion
kable(apply(full_dat, 2, num_missing)/nrow(dat))

# look at races with missingness, choosing arbitrary age present in all races
# and first weather measurement
full_dat%>%filter(is.na(TD))%>%group_by(Race)%>%filter(Age==25)
## complete case analysis
full_dat<-full_dat[complete.cases(full_dat),]
## create table 1
tbl_summary(select(full_dat, c(Age, Sex, Percent_CR, Time, Race)), by=Race, missing="no",
            statistic = list(all_continuous()~ "{mean} ({sd})",
                             all_categorical()~ "{n} ({p}%)"))%>%
  modify_header(all_stat_cols() ~ "**{level}**  \nN = {n}")%>%
  modify_caption("Runner Characteristics by Race")%>%
  as_kable_extra(booktabs=TRUE)
tbl_summary(select(full_dat, -c(Year, CR, cr.time, Sex, Age, Percent_CR)), by=Race, missing="no",
            statistic = list(all_continuous()~ "{mean} ({sd})",
                             all_categorical()~ "{n} ({p}%)"))%>%
  modify_header(all_stat_cols() ~ "**{level}**  \nN = {n}")%>%
  modify_caption("Weather Characteristics by Race")%>%
  as_kable_extra(booktabs=TRUE)
# data frames for slope
a<-data.frame(x=seq(14, 25, 1))%>%mutate(y=200-9*x)
b<-data.frame(x=seq(25, 50, 1))%>%mutate(y=-50+1*x)
c<-data.frame(x=seq(50, 75, 1))%>%mutate(y=-200+4*x)
d<-data.frame(x=seq(75, 95, 1))%>%mutate(y=-500+8*x)

#### plot percent course record with smoothed line, by sex and race
g<-ggplot(full_dat, aes(x=Age, y=Percent_CR))+
  geom_point(alpha=.5, color="lightblue")+
  geom_line(data=a, aes(x, y, color="red"))+
  annotate("text", 17, 0, label="slope=-9", color="red", size=3)+
  geom_line(data=b, aes(x, y, color="red"))+
  annotate("text", 37, -30, label="slope=1", color="red", size=3)+
  geom_line(data=c, aes(x, y, color="red"))+
  annotate("text", 65, 25, label="slope=4", color="red", size=3)+
  geom_line(data=d, aes(x, y, color="red"))+
  annotate("text", 87, 160, label="slope=8", color="red", size=3)+
  geom_smooth(aes(x=Age, y=Percent_CR, color="blue"), dat=full_dat[full_dat$Sex=="M",])+
  geom_smooth(aes(x=Age, y=Percent_CR, color="purple"), dat=full_dat[full_dat$Sex=="F",])+
  #facet_wrap(~Race, )+
  scale_color_manual(values=c("blue","purple", "red"), labels=c("Male", "Female", "Approximate Slope"))+
  labs(x="Age", y="Percent CR", title=
         "Figure 1: Finishing Times as Percent Course Record by Sex and Race")+
  theme(legend.position = "top",
        panel.background = element_rect(fill="white"),
        strip.background =element_rect(fill="grey90", linewidth = .2,
```

```
                                        colour = "grey70"),
        panel.grid.major = element_line(color="grey70", linewidth=.2),
        panel.grid.minor = element_line(color="grey70", linewidth=.2),
        legend.key = element_rect(fill = "grey90")
        )
g
# create bins for age
full_dat<-full_dat%>%mutate(Age_cat=case_when(Age<20~"<20",
                                              Age<50~"20-49",
                                              Age<75~"50-74",
                                              TRUE~"75+"))

### plot times by race and sex
ggplot(full_dat, aes(x=Percent_CR, fill=Sex))+
  geom_density(alpha=.75)+
  facet_wrap(~Age_cat)+
  scale_fill_manual(values=c("blue","purple"), labels=c("Female", "Male"))+
  labs(x="Percent CR", y="Density", title="Figure 2: Distribution of Finishing Times by Sex and Age Cat
  theme(legend.position = "top",
        panel.background = element_rect(fill="white"),
        strip.background =element_rect(fill="grey90", linewidth = .2,
                                        colour = "grey70"),
        panel.grid.major = element_line(color="grey70", linewidth=.2),
        panel.grid.minor = element_line(color="grey70", linewidth=.2),
        legend.key = element_rect(fill = "grey90")
        )
trim_dat<-full_dat[complete.cases(full_dat),]
trim_dat$Flag<-factor(trim_dat$Flag, levels=c("White", "Green", "Yellow", "Red"))
## table of data characteristics by flag
tbl_summary(select(trim_dat, c(Race, Sex, Age, Percent_CR,Time, Flag)), by=Flag,
            missing="no",
            statistic = list(all_continuous()~ "{mean} ({sd})",
                             all_categorical()~ "{n} ({p}%)"))%>%
  add_p()%>%
  modify_header(all_stat_cols() ~ "**{level}**  \nN = {n}")%>%
  modify_caption("Data Characteristics by Flag")%>%
  as_kable_extra(booktabs=TRUE)
ggplot(full_dat[!is.na(full_dat$Flag),], aes(x=Percent_CR, color=Flag))+
  geom_density(alpha=.75, linewidth=.75)+
  facet_wrap(~Age_cat)+
  scale_color_manual(values=c("green", "red","black", "gold"),
                     labels=c("Green", "Red", "White", "Yellow"))+
  labs(x="Percent_CR", y="Density", title=
         "Figure 3: Distribution of Finishing Times by Age and Flag")+
  theme(legend.position = "top",
        panel.background = element_rect(fill="white"),
        strip.background =element_rect(fill="grey90", linewidth = .2,
                                        colour = "grey70"),
        panel.grid.major = element_line(color="grey70", linewidth=.2),
        panel.grid.minor = element_line(color="grey70", linewidth=.2),
        legend.key = element_rect(fill = "grey90")
        )
```

```r
p<-trim_dat%>%group_by(Age_cat, Sex, PM2.5)%>%summarize(PCR=mean(Percent_CR))
o<-trim_dat%>%group_by(Age_cat, Sex, Ozone)%>%summarize(PCR=mean(Percent_CR))

aqi_dat<-data.frame(Age_cat=c(p$Age_cat, o$Age_cat), Sex=c(p$Sex, o$Sex), Val=c(p$PM2.5, o$Ozone),
                    Type=c(rep("PM2.5", nrow(p)), rep("Ozone", nrow(o))), PCR=c(p$PCR, o$PCR))

aqi_plot<-ggplot(data=aqi_dat, aes(x=Val, y=PCR, color=Age_cat, linetype=Sex))+
  geom_smooth(se=FALSE)+
  scale_color_manual(values=c("blue", "purple", "orange", "red"),
                     labels=c("<20", "20-49", "50-74", "75+"))+
  scale_linetype_manual(values=c("solid", "dotted"),
                        labels=c("M", "F"))+
  labs(x="AQI Value", y="Average Percent Course Record", title=
         "Figure 4: Effect of AQI Measures on Finishing Times")+
  theme(legend.position = "top",
        panel.background = element_rect(fill="white"),
        strip.background =element_rect(fill="grey90", linewidth = .2,
                                       colour = "grey70"),
        panel.grid.major = element_line(color="grey70", linewidth=.2),
        panel.grid.minor = element_line(color="grey70", linewidth=.2),
        legend.key = element_rect(fill = "grey90")
        )+
  facet_wrap(~Type)

aqi_plot

# linear regression of time with WBGT interactions
# set baseline age group to 20-49
trim_dat$Age_cat<-factor(trim_dat$Age_cat, ordered = FALSE)
trim_dat$Age_cat<-relevel(trim_dat$Age_cat, ref="20-49")

l_f<-lm(Time~Age_cat*WBGT+Age_cat*Ozone+Age_cat*PM2.5, data=trim_dat[trim_dat$Sex=="F",])
l_m<-lm(Time~Age_cat*WBGT+Age_cat*Ozone+Age_cat*PM2.5, data=trim_dat[trim_dat$Sex=="M",])

# summary table of regression coefficients
tbl_regression(l_f)%>%
  modify_caption("Linear Regression of Time (m) with WBGT Interactions in Females")%>%
  as_kable_extra(booktabs=TRUE)

tbl_regression(l_m)%>%
  modify_caption("Linear Regression of Time (m) with WBGT Interactions in Males")%>%
  as_kable_extra(booktabs=TRUE)
# Create correlation matrix
cor_mat <- trim_dat%>%
  select(c(RH, SR, DP, Wind, WBGT, Ozone, PM2.5)) %>%
  cor(use = "complete.obs")

# Plot correlation matrix
corrplot(cor_mat, method = "square", type = "lower", diag = FALSE,mar=c(0,0,2,0),
         tl.cex = 0.7, tl.col = "black", tl.srt = 30, title = " Figure 5. Correlation Between Weather Va
q=trim_dat%>%group_by(RH)%>%summarize(time=mean(Time))
plot(q$RH, q$time)
plot(y=trim_dat$Time, trim_dat$RH)
```

```r
plot(y=trim_dat$Time, trim_dat$SR)
plot(y=trim_dat$Time, trim_dat$DP)
plot(y=trim_dat$Time, trim_dat$Wind)
plot(y=trim_dat$Time, trim_dat$WBGT)
plot(y=trim_dat$Time, trim_dat$Ozone)
plot(y=trim_dat$Time, trim_dat$PM2.5)
# fit linear model for weather conditions
weather_fit<-lm(Time~RH+SR+DP+Wind+WBGT+Ozone+PM2.5+Age+I(Age^2), data=trim_dat[trim_dat$Sex=="F",])
weather_fit2<-lm(Time~RH+SR+DP+Wind+WBGT+Ozone+PM2.5+Age+I(Age^2), data=trim_dat[trim_dat$Sex=="M",])

# summary table of regression coefficients
tbl_regression(weather_fit, intercept=TRUE)%>%
  modify_caption("Linear Regression of Time (m) and Weather Conditions in Females")%>%
  as_kable_extra(booktabs=TRUE)

tbl_regression(weather_fit2, intercept=TRUE)%>%
  modify_caption("Linear Regression of Time (m) and Weather Conditions in Males")%>%
  as_kable_extra(booktabs=TRUE)
```