Project-1

Zhaoxiang Ding

Exploratory Analysis

There are 11564 records with 14 variables in the data set. Each row represent the fastest finishing time among men or women at a year of age, compared with the course record (`%CR`) among Boston, Chicago, New York, Twin Cities, Grandma's Marathons for 17 to 24 years, 1993-2016, and the corresponding weather conditions.

The participants are range from age 14 to 91 for men, and 14 to 88 for women. The distribution of the weather conditions (represented by `Flag`, a variable that indicates the wet bulb globe temperature(WBGT)) is shown in Table 1. `WBGT` is the Weighted average of dry bulb, wet bulb, and globe temperature (measured in celsius), which measure the heat stress considering temperature, humidity and solar radiation. Therefor, `Flag` is a reasonable proxy for the weather conditions of marathons, with `White` indicating WBGT is lower than 10 degrees, `Green` indicating WBGT is between 10 and 18 degrees, `Yellow` indicating WBGT is between 18 and 23 degrees, and `Red` indicating WBGT is between 20 and 28 degrees, and `Black` indicating WBGT is higher than 28 degrees. Table 1 shows that the distribution of weather condition is not balanced, with more than 40% of records are in `Green` flag and only 5% of records are in `Red` flag. No races were held in `Black` flag condition. There are 491 records (4%) missing the `Flag` information, as long as other weather conditions, which is because the race was canceled due to weather conditions. However, these records still contains the information of the fastest finishing time. The deatils of the races with missing weather conditions are: Chicago Marathon, New York City Marathon, Twin Cities Marathon in 2011 and Grandma's Marathon in 2012. Excluding those records from our analysis may bring bias as it's anticipated that the weather conditions may have an impact on the marathon performance. However, considering the relative small portion of missingness, we will exclude them from the analysis.

Table 1: Summary of Flag variable in the dataset

Flag	Count
	491
Green	4706
Red	592

White	3753
Yellow	2022

This report aim to examine 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, identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance and evaluate the impact of air quality on marathon performance. The analysis is carried out in R (version 4.4.0) using `ggplot2`, `dplyr`, `lubridate`, `kableExtra`, `gtsummary`, `stringr`, `ggpubr`, `RColorBrewer`, `gridExtra` and `latex2exp` packages.

Examine effects of increasing age on marathon performance in men and women

It is proposed that aging will have a negative impact on marathon performance as older individuals will have reduced ability to tolerate heat stress. However, does this assumption translate to decreased running performance in hot environments and does the effect of aging differ between men and women remains unknown. We compared the fastest finish time against age between men and women. Here, the fastest finish time are calculated by $CR + \%CR * CR$, where CR is the course record of the race in the given year for gender and $\%CR$ is the Percent off current course record for gender. Since different races may take different time to finish, we compared the performance among different races across different years by calculating smoothed conditional means. Our result (Figure 1) shows that both men and women have a U-shaped relationship between age and fastest finish time, with women have a higher finish time compare to men across all ages. The fastest finish time first drop as age increase until 26 years old, then increase as age increase, for both men and women. It need to be noticed that the slope of the curve changed after 52 years old, with the fastest finish time increase faster as age increase. The slope of the curve is roughly the same between men and women before 75 years old, but the slope varies among different races after 75 years old. The result indicates that aging have a negative impact on marathon performance, with the impact is more significant after 52 years old. Eventhough the fastest finish time drop as age increase until 26 years old, considering the fact that body function is not going to drop at such a young age, this may only due to the fact that young runner may not have enough experience to run a marathon.

The result also suggest that the effect of aging on marathon performance is consistent between gender until 75 years old. But the effect after age 75 varies across races. The effect decrase after 75 years old for both gender in New york City Marathon, and stayed the same for chicago marathon. While in Boston Marathon, the effect only decrease among women, and in twin cities marathon, only among men. In Grandma's marathon, the effect decrease for men and first decrease, then increase among women.

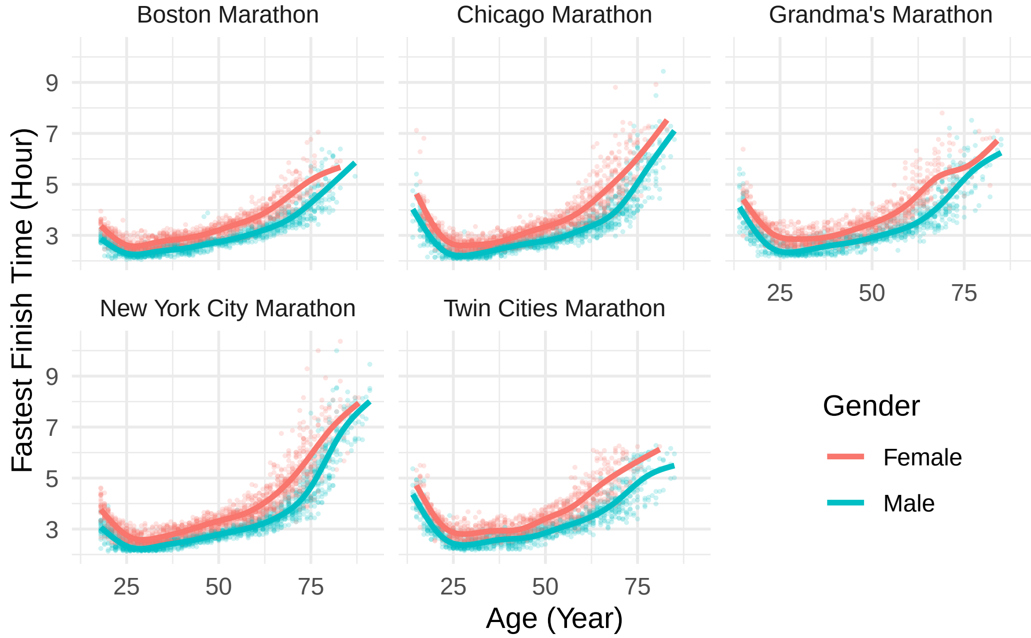

Figure 1: Scatter plot of marathon performance against age among different races

Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

In order to examine the effect of environmental conditions on marathon performance, we first define what should be the variable to represent the marathon performance. The fastest finish time can serve as a good proxy for marathon performance, but as described earlier and shown in Figure 1, the fastest finish time differs among different races. To make the comparison fair, we will utilize the percentage off record, $\%CR$ to represent the marathon performance as the difference of $\%CR$ should remain the same among different races condition on all other factors remain the same. We compare the smoothed conditional means of $\%CR$ against age among different weather conditions. The result (Figure 2) shows similar pattern as the fastest finish time, with the $\%CR$ first drop as age increase until 26 years old, then increase as age increase. The smoothed conditional means is roughly the same among different weather conditions, until 75 years old, where slope of the mean start to decrease whil the others remain the same. But this can not gurantee that the effect of aging start to be less effective when the weather temperature is high as there are only little data points collected for the senior persons (represented by the scatter dots). A slightly bigger difference among different weather conditions can be observed among runner age between 52 and 75 years old, but in general, the effect of weather conditions on marathon performance is not significant.

It can be shown that among the times that the record is broken, the percentage of `Green` flag

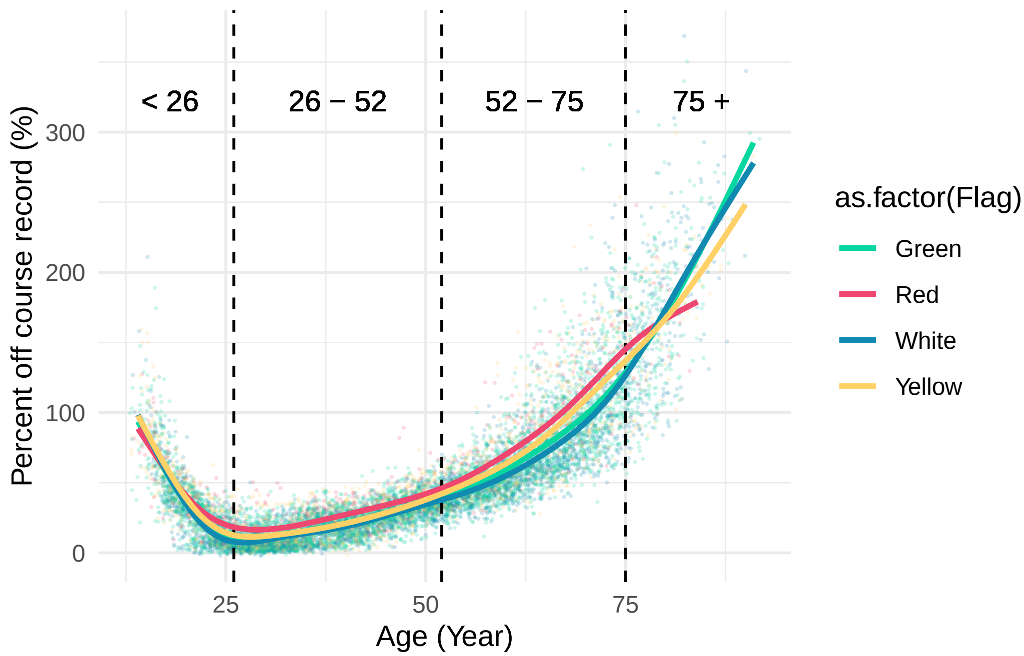

Figure 2: Scatter plot of marathon performance against age among different weather conditions

is higher than the percentage of `Green` flag in the data by more than 10 percentage points and the percentage of `Yellow` flag is lower than the percentage of `Yellow` flag in the data by more than 10 percentage points too.

Since we have already described the relationship between age and marathon performance that in general, marathon performance decrease as age increase. Defining a new way to represent marathon performance, which will show no difference among age is crucial so that we can compare the effect of weather conditions on marathon performance. We will use $\frac{\%CR - E[\%CR|Age]}{E[\%CR|Age]}$, the normalized difference between the $\%CR$ and the expectation of $\%CR$ given the age, to represent the marathon performance of each rows, where $E[\%CR|Age]$ is the expectation of $\%CR$ given the age. This new variable is the percentage of $\%CR$ that is off the expectation of $\%CR$ given the age, and it is expected to be 0 across all ages if the weather conditions have no impact on marathon performance. A natural way to estimate $E[\%CR|Age]$ is by calculating the means, stratified by age. However, as shown in Table 1, the distribution of weather conditions is not balanced. If the assumption that weather do have effect on marathon performance is true, then the arithmetic mean will be biased. To address this issue, we will use the weighted mean to estimate $E[\%CR|Age]$, where the weight is the proportion of each weather condition in the data set. With this new variable, we can compare the marathon performance among different weather conditions, and examine whether the impact of weather conditions on marathon performance differs across age and sex. In Figure 3, we separate the records into 2 groups, below average marathon performance and above average marathon performance, corresponding to whether the normalized difference is below or above 0. We also compare the difference of using weighted mean and arithmetic mean to estimate $E[\%CR|Age]$. The result shows that the temperature of 4 different measurements are all higher in those records with above average marathon performance. It also shows that there is little difference between using weighted mean and arithmetic mean to estimate $E[\%CR|Age]$. Nonetheless, we still adopt weighted mean to estimate $E[\%CR|Age]$ because it can avoid bias by theory.

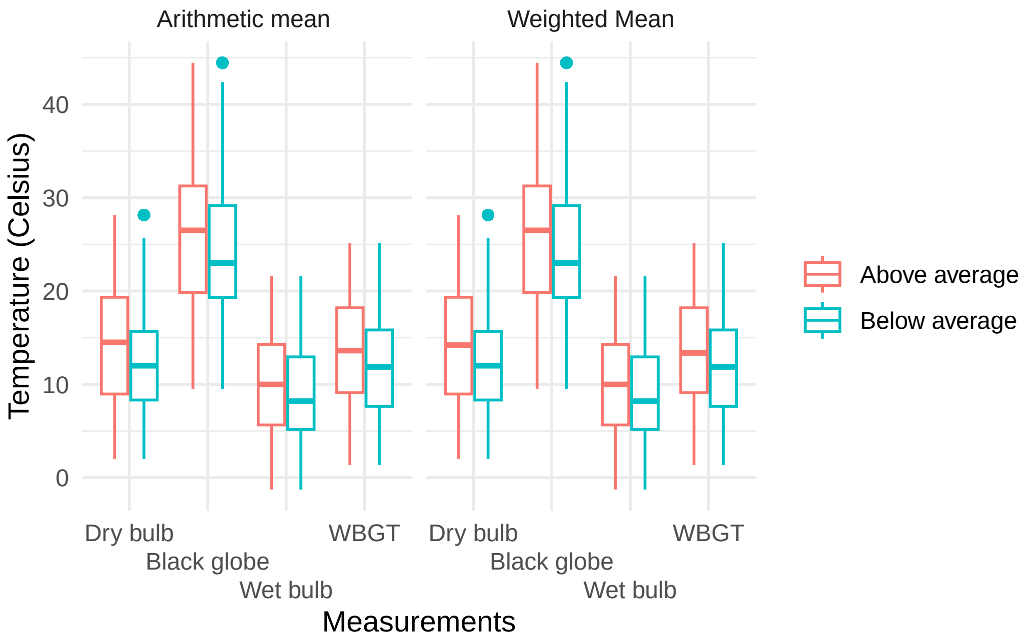

Figure 3: Boxplot of temperatures measured in different equipment between above and below average marathon performance records.

Table 2 shows that the mean of all weather conditions are significantly different between above and below average marathon performance records, except for the Percent relative humidity. The average of temperature, solar radiation and dew point are all higher in the records with above average marathon performance, while the average of wind speed is lower in the records with above average marathon performance. The result indicates that the weather conditions do have an impact on marathon performance, with higher temperature, solar radiation and dew point and lower wind speed are associated with better marathon performance.

Table 2: Siginificant test of the difference of the average weather conditions between above and below average marathon performance records

Characteristic	**Above average** N = 5,386[1]	**Below average** N = 5,687[1]	**p-value**[2]
Flag			<0.001

White	1,552 (29%)	2,201 (39%)	
Green	2,263 (42%)	2,443 (43%)	
Yellow	1,162 (22%)	860 (15%)	
Red	409 (7.6%)	183 (3.2%)	
Td..C	14.2 (9.0, 19.3)	12.0 (8.3, 15.7)	<0.001
Tw..C	10.0 (5.7, 14.3)	8.2 (5.2, 12.9)	<0.001
Tg..C	26 (20, 31)	23 (19, 29)	<0.001
WBGT	13.4 (9.1, 18.2)	11.9 (7.6, 15.8)	<0.001
X.rh	53 (1, 64)	52 (1, 64)	0.5
SR.W.m2	513 (390, 627)	513 (354, 602)	<0.001
DP	7 (2, 12)	4 (0, 10)	<0.001
Wind	9.8 (7.0, 11.7)	10.0 (7.6, 12.2)	<0.001

[1]n (%); Median (Q1, Q3)
[2]Pearson's Chi-squared test; Wilcoxon rank sum test

To determine whether the effect of weather conditions on marathon performance differs across age and sex, we stratify the records by `flag` to represent for different weather conditions. As shown in Figure 4, runner are more likely to have bad performance (positive normalized difference) when WBGT is higher than 23 degrees (Red flag), especially for runner between 20 to 30 years old, where the normalized difference reaches it's peak to 0.5 (50% worse than average performance). After 30 years old, the normalized difference decrease as age increase, and Female runnner experienced a faster decrease than male runner, indicating that the effect of weather conditions on marathon performance is weaker in female runners than male runner after 30 years old. The result also shows that runners performance is roughly the same as the average performance when WBGT is between 10-18 degrees (Green flag). The performance ise slightly worse when WBGT is between 18-23 degrees (Yellow flag) and better when WBGT is lower than 10 degrees (White flag).

By stratifying the records by age group, we can see that the effect of weather conditions on marathon performance is more significant among older runners. The age group is defined as `< 26`, `26-52`, `52-75`, `75 +`. The groups are selected by different slopes based on Figure 2. The result (Figure 5) shows a similar result with Figure 4, with the effect of weather conditions on marathon performance is more significant among runners between 26 and 52 years old, male runners experience a stronger effect by weather conditions and the effect is weaker among older runners.

Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

The normalized difference we defined earlier proved to be useful to examine the effect of weather conditions on marathon performance while excluding the effect of aging. In order to

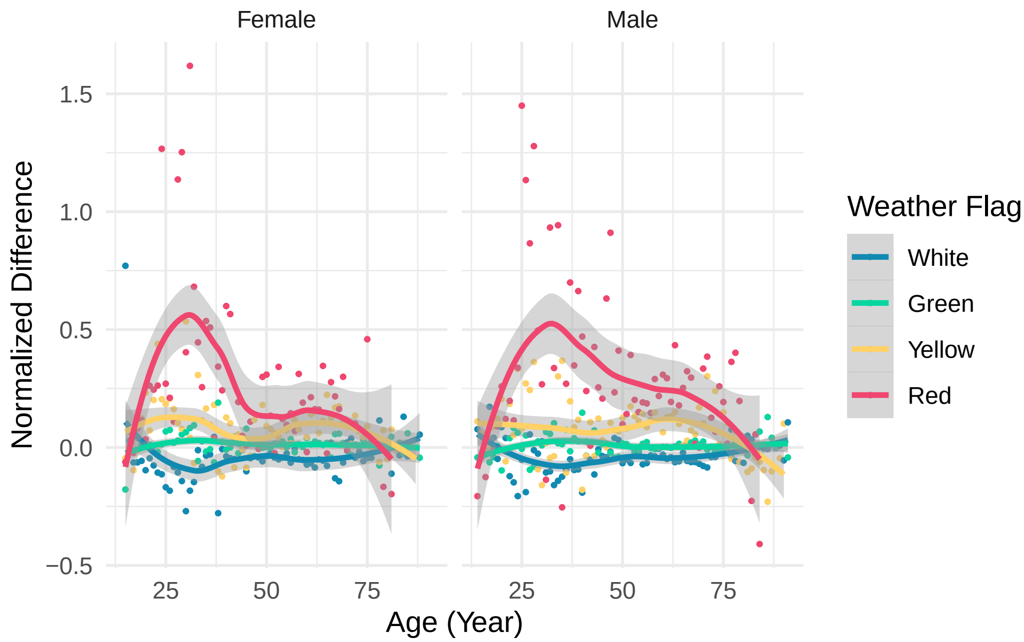

Figure 4: Scatter plot of normalized difference against age among different weather conditions

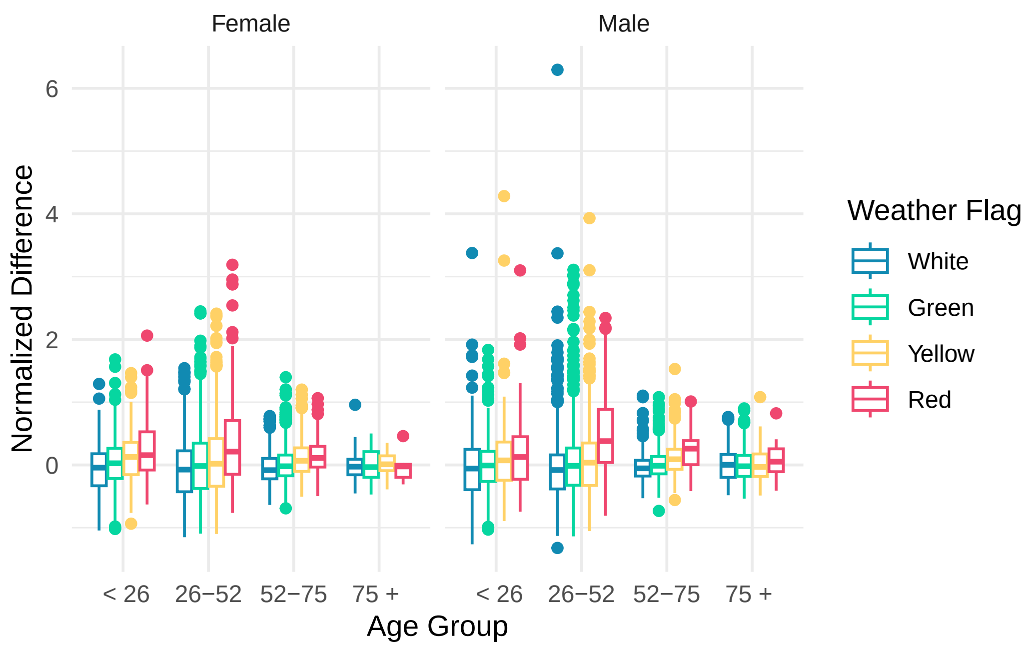

Figure 5: Boxplot of normalized difference against age group among different weather conditions

determine the largest impact on marathon performance, we calculated the correlation between the normalized difference and the weather parameters who are significant in Table 2. The result (Figure 6) shows that all the temperatures (Dew point, Dry bulb, Wet bulb, etc.) are share a similar pattern or correlation. The correlation first increase until around age 60, then start to decrease. The correlation between the normalized difference and the wind is the exact opposite, with the correlation first decrease until around age 60, then start to increase. The correlation between the normalized difference and the solar radiation is the lowest among all weather parameters, with the correlation is roughly the same across all ages. The correlation measures the strength of the linear relationship between the weather parameters and marathon performance, a positive relation indicates that the increase of weather parameter is associated with worse marathon performance (a higher normalized difference) and vice versa. The magnitude of the correlation can be interpreted as how likely will the marathon performance be affected by the weather parameter. The result indicates that the marathon performance is more likely to be affected by the temperature (adjusted by other weather conditions or unadjusted) than any other weather parameters, and older runner is more likely to be affected by the weather conditions than younger runner.

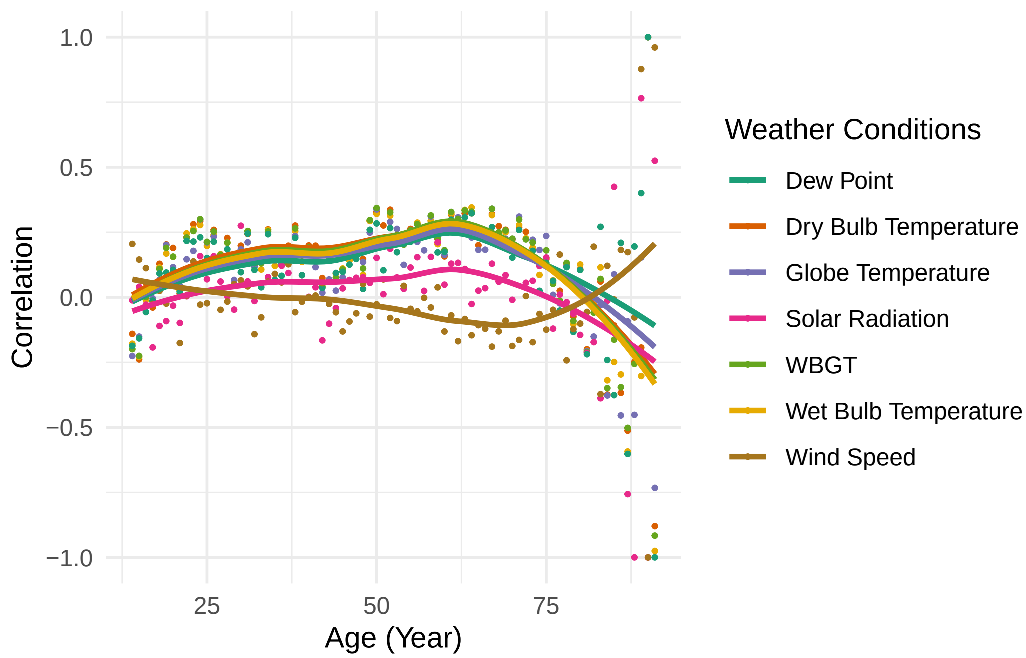

Figure 6: Scatter plot of correlation between weather parameters and normalized difference against age

The result which stratified the records by age group (Figure 7) shows a similar pattern as Figure 6. It also shows that the variance of correlation is higher among runner above 75 years old. Indicating the decreasing effect of weather conditions on marathon performance among older runner shown in Figure 6 may be unreliable.

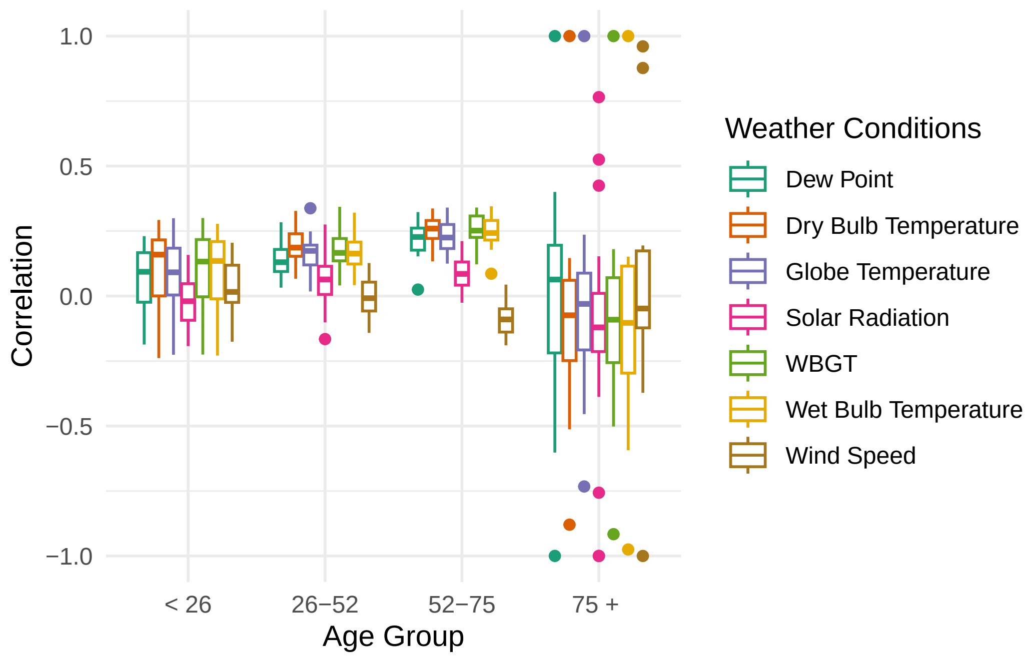

Figure 7: Boxplot of correlation between weather parameters and normalized difference against age group

Evaluate the impact of air quality on marathon performance

Except for the weather conditions, air quality may also have an effect on marathon performance. We will explore the impact of air quality on marathon performance by examining the correlation between the air quality index (AQI) and the normalized difference. The AQI of each race is calculated by averaging the AQI of all hours measured by 8-HR RUN AVERAEG BEGIN HOUR. The data set provides other kinds of measurements of air quality, but only record measured by 8-HR RUN AVERAEG BEGIN HOUR has no missing value and is measured in the same way across all locations and years.

The result (Figure 8) shows that the correlation between AQI and the normalized difference is roughly the same across all ages, with the average correlation is only 0.05. To further explore the effect of air quality, we also fit a linear model to examine the effect of other weather parameters on marathon performance, condition on air quality. The model is fitted by $Diff \sim \beta_0 + \beta_1 X + \beta_2 * AQI + \beta_3 * X * AQI$, where $Diff$ is the normalized difference, X is the weather parameter, and AQI is the air quality index. The estimated coefficient of β_1 and β_3 , and the corresponding p-value are reported in Table 3. The result shows that, only the effect of Solar Radiation is not siginificantly associated with air quality, and after condition on air quality, the effect of relative humidity is still not significant. The estimated value of β_3 for other weather parameters are all significantly above 0, indicating that the effect of weather parameters on marathon performance is stronger when AQI is higher (worse air quality).

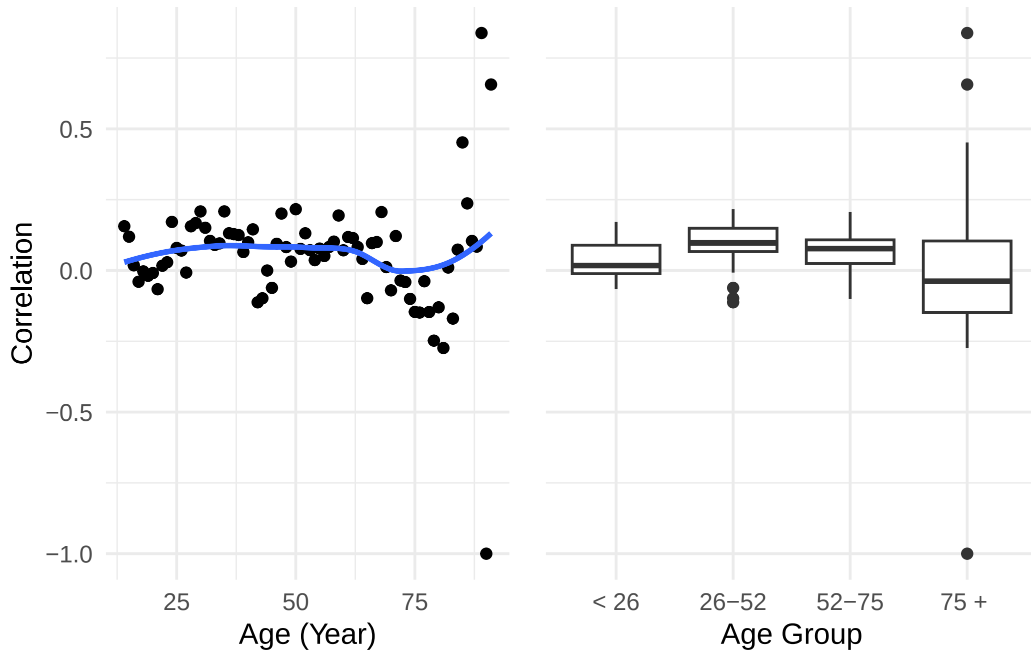

Figure 8: Scatter plot and box plot of correlation between air quality index and normalized difference against age

Table 3: Estimated coefficient and p-value of weather parameters on normalized difference, condition on air quality

Parameter	Estimate	P-value
Dry bulb temperature	0.009	< 0.001*
Dry bulb temperature * Aqi	< 0.001	0.001*
Wet bulb temperature	-0.004	0.116
Wet bulb temperature * Aqi	0.001	< 0.001*
Black globe temperature	0.006	< 0.001*
Black globe temperature * Aqi	< 0.001	0.011*
Wet Bulb Globe Temperature	0.002	0.252
Wet Bulb Globe Temperature * Aqi	< 0.001	< 0.001*
Solar Radiation	< 0.001	0.012*
Solar Radiation * Aqi	< 0.001	0.146
Dew Point	-0.009	< 0.001*
Dew Point * Aqi	0.001	< 0.001*
Wind Speed	0.004	0.113

Wind Speed * Aqi	< 0.001	0.002*
Relative Humidity	< 0.001	0.477
Relative Humidity * Aqi	< 0.001	0.089

Conclusion

Our analysis shows that aging have an overall negative impact on marathon performance, in terms of fastest finish time and percentage off course record. All weather conditions except for the relative humidity have a significant impact on marathon performance, in terms of normalized difference, with higher temperature, solar radiation and dew point and lower wind speed are associated with worse marathon performance. Runners around 26 years old are the fastest runner compare to all other age, and effect of aging start to accelerate around 53 years old (Figure 1 and Figure 2).Wet Bulb Globe Temperature only have effect on marathon performance when the temperature is over 23 degrees (Red Flag). The effect of Wet Bulb Globe Temperature varies across age and sex, where the effect is stronger among younger runners (between 20 to 30 years old) and among male runners (Figure 4). Even though young runners suffered the biggest effect of bad weather conditions, old runners are more likely to be affected by the weather conditions (Figure 6). Temperatures measured in all kinds of way(Dry bulb, wet bulb, etc.) are the mostly related to marathon performance.

Code Appendix

```
# load libraries
library(dplyr)
library(tidyr)
library(kableExtra)
library(gtsummary)
library(lubridate)
library(ggplot2)
library(stringr)
library(latex2exp)
library(ggpubr)
library(RColorBrewer)
library(gridExtra)
# Read data
data <- read.csv("../Data/project1.csv")
air <- read.csv("../Data/aqi_values.csv")
# change coloumn name to make it easier to understand
colnames(data)[[1]] <- 'Race'
```

```r
colnames(data)[[3]] <- 'Sex'

# Table 1
table(data$Flag) %>%
  kable(col.names = c('Flag', 'Count')) %>%
  kable_styling()
# Missing record summary

data_na <- data[data$Flag == '',]
data <- data[data$Flag != '',]

data_na$location <- case_when(data_na$Race == 0 ~ 'Boston Marathon',
                              data_na$Race == 1 ~ 'Chicago Marathon',
                              data_na$Race == 2 ~ 'New York City Marathon',
                              data_na$Race == 3 ~ 'Twin Cities Marathon',
                              data_na$Race == 4 ~ 'Grandma's Marathon')
data_na$marathon <- paste(data_na$location, data_na$Year, sep = ' in ')
na_df <- unique(data_na$marathon)
na_df <- str_split(na_df, ' in ')
na_df <- as.data.frame(do.call(rbind, na_df))
colnames(na_df) <- c('Race', 'Year')
# na_df %>%
#   kable() %>%
#   kable_styling()
# Course record data
record <- read.csv("../Data/course_record.csv")
record$Race <- case_when(
  record$Race == 'B' ~ 0,
  record$Race == 'C' ~ 1,
  record$Race == 'NY' ~ 2,
  record$Race == 'TC' ~ 3,
  record$Race == 'D' ~ 4
)

colnames(record)[[4]] <- 'Sex'
record$seconds <- period_to_seconds(hms(record$CR))
record$Sex <- ifelse(record$Sex == 'M', 1, 0)

# Merge data
data_record <- left_join(data, record, by = c('Race', 'Year', 'Sex'))

# fastest running time
```

```r
data_record$Time <- data_record$seconds + (data_record$X.CR/100) * data_record$seconds
data_record$locations <- case_when(data_record$Race == 0 ~ 'Boston Marathon',
                                   data_record$Race == 1 ~ 'Chicago Marathon',
                                   data_record$Race == 2 ~ 'New York City Marathon',
                                   data_record$Race == 3 ~ 'Twin Cities Marathon',
                                   data_record$Race == 4 ~ 'Grandma’s Marathon')
# Fig 1
ggplot(data_record, aes(x = Age..yr., y = Time, color = as.factor(Sex))) +
  geom_point(size = 0.2, alpha = 0.2) +
  geom_smooth(se = F) +
  scale_color_discrete(name = 'Gender', label = c('Female','Male'))+
  scale_y_continuous(breaks = c(10800,18000, 25200, 32400),
                     labels = c(3,5,7,9),
                     name = 'Fastest Finish Time (Hour)') +
  scale_x_continuous(name = 'Age (Year)')+
  facet_wrap(~locations) +
  theme_minimal() +
  theme(legend.position = c(.85,.2))

# Figure 2
ggplot(data_record, aes(x = Age..yr., y = X.CR)) +
  geom_point(aes(color = as.factor(Flag)),
             size = 0.05, alpha = 0.2, position = position_jitter(width = 1)) +
  geom_vline(xintercept = c(26, 52, 75), linetype = 'dashed') +
  geom_text(x = 18, y = 300, label = '< 26', vjust = -1) +
  geom_text(x = 39, y = 300, label = '26 - 52', vjust = -1) +
  geom_text(x = 63.6, y = 300, label = '52 - 75', vjust = -1) +
  geom_text(x = 85, y = 300, label = '75 + ', vjust = -1) +
  geom_smooth(aes(color = as.factor(Flag)), se = F) +
  scale_color_manual(values = c('#06d6a0', '#ef476f', '#118ab2', '#ffd166')) +
  scale_y_continuous(name = 'Percent off course record (%)') +
  scale_x_continuous(name = 'Age (Year)') +
  theme_minimal()

# Record breaking summary
# won't be shown in the final report
record_flag <- left_join(record, data[,1:4], by = c('Race', 'Year', 'Sex'))
record_flag <- record_flag[!is.na(record_flag$Flag),]
record_break <- record_flag %>%
  distinct(CR, .keep_all = T)

flag_break_record <- round(as.numeric(summary(as.factor(record_break$Flag))*100/nrow(record_
```

```
flag_data <- round(as.numeric(summary(as.factor(data$Flag))*100/nrow(data)),3)

flag_df <- data.frame(Flag = c('Green', 'Red', 'White', 'Yellow'),
                      Record = flag_break_record,
                      Data = flag_data)
# flag_df %>%
#   kable(col.names = c('Flag', 'Percentage in record breaking cases (%)', 'Percentage in Da
#   kable_styling()
# calculate mean and weighted mean

Avg_CR_unweighted <- data_record %>% #unweighted mean
  group_by(Sex, Age..yr.) %>%
  summarise(mean_CR_unweight = mean(X.CR))

Flag_Weight <- data_record %>% # calculate weight based on flags
  group_by(Flag) %>%
  summarise(weight = n()/nrow(data_record))

data_record_weight <- left_join(data_record, Flag_Weight, by = 'Flag')

# weighted mean
Avg_CR_weighted <- data_record_weight %>%
  group_by(Sex, Age..yr.) %>%
  summarise(mean_CR_weighted = weighted.mean(X.CR, w = weight),
            mean_TdC_weighted = weighted.mean(Td..C, w = weight),
            mean_TwC_weighted = weighted.mean(Tw..C, w = weight),
            mean_TgC_weighted = weighted.mean(Tg..C, w = weight),
            mean_WBGT_weighted = weighted.mean(WBGT, w = weight),
            mean_SR.W.m2_weighted = weighted.mean(SR.W.m2, w = weight),
            mean_DP_weighted = weighted.mean(DP, w = weight),
            mean_Wind_weighted = weighted.mean(Wind, w = weight))

# merge data
data_record_weight <- left_join(data_record_weight,
                                Avg_CR_weighted,
                                by = c('Sex', 'Age..yr.'))
data_record_weight <- left_join(data_record_weight,
                                Avg_CR_unweighted,
                                by = c('Sex', 'Age..yr.'))

data_record_weight$below_avg <- ifelse(data_record_weight$X.CR < data_record_weight$mean_CR_u
data_record_weight$below_avg_w <- ifelse(data_record_weight$X.CR < data_record_weight$mean_CR
```

```
#converge data to long format for Fig 3
data_weight_long <- pivot_longer(data_record_weight,
                                 cols = c('Td..C', 'Tw..C', 'Tg..C', 'WBGT'),
                                 names_to = 'Type', values_to = 'Temp')

data_weight_long <- pivot_longer(data_weight_long,
                                 cols = c('below_avg', 'below_avg_w'),
                                 names_to = 'Cr_type',
                                 values_to = 'Below_avg_CR')
data_weight_long$Cr_type <- ifelse(data_weight_long$Cr_type == 'below_avg',
                                   'Arithmetic mean ',
                                   'Weighted Mean')
data_weight_long$Below_avg_CR <- ifelse(data_weight_long$Below_avg_CR == 1,
                                        'Below average',
                                        'Above average')

# Fig3
ggplot(data_weight_long, aes(x = Type, y = Temp, color = as.factor(Below_avg_CR))) +
  geom_boxplot() +
  scale_color_discrete(name = '') +
  scale_x_discrete(name = 'Measurements',
                   labels = c('Dry bulb', 'Black globe', 'Wet bulb', 'WBGT'),
                   guide = guide_axis(n.dodge=3)) +
  scale_y_continuous(name = 'Temperature (Celsius)') +
  facet_wrap(~Cr_type) +
  theme_minimal()

# normalized difference
data_record_weight$Avg_Cr_Diff <- (data_record_weight$X.CR -
                                     data_record_weight$mean_CR_weighted)/
  data_record_weight$mean_CR_weighted

data_record_weight$Flag <- factor(data_record_weight$Flag,
                                  ordered = T,
                                  levels = c('White','Green', 'Yellow', 'Red'))
data_record_weight$below_avg_w <- ifelse(data_record_weight$below_avg_w == 1,
                                         'Below average',
                                         'Above average')

# table 2
tbl_summary(data_record_weight, by = below_avg_w,
            include = c(Flag, Td..C, Tw..C, Tg..C, WBGT, X.rh, SR.W.m2, DP, Wind)) %>%
```

```
  add_p()
# Tg flag
# not used in the final report
data_record_weight$Tg_flag <- case_when(data_record_weight$Tg..C < 19.475000 ~ 1,
                                        data_record_weight$Tg..C < 24.955556 ~ 2,
                                        data_record_weight$Tg..C < 30.000000 ~ 3,
                                        TRUE ~ 4)
data_record_weight$Tg_flag <- as.factor(data_record_weight$Tg_flag)

data_record_weight$Avg_Cr_Diff <- (data_record_weight$X.CR -
                                      data_record_weight$mean_CR_weighted)/
  data_record_weight$mean_CR_weighted

data_record_weight$Sex <- ifelse(data_record_weight$Sex == 1, 'Male', 'Female')

# Fig 4
data_record_weight %>%
  group_by(Sex, Age..yr., Flag) %>%
  summarise(mean_cr = mean(Avg_Cr_Diff)) %>%
  ggplot(aes(x = Age..yr., y = mean_cr, color = as.factor(Flag))) +
  geom_point(size= 0.5) +
  geom_smooth(se = T) +
  scale_color_manual(values = c('#118ab2','#06d6a0', '#ffd166', '#ef476f'),
                     name = 'Weather Flag') +
  scale_y_continuous(name = 'Normalized Difference') +
  scale_x_continuous(name = 'Age (Year)') +
  facet_wrap(~Sex)+
  theme_minimal()
# Age group
data_record_weight$Age_grp <- case_when(data_record_weight$Age..yr. < 26 ~ ' < 26',
                                        data_record_weight$Age..yr. < 52 ~ '26-52',
                                        data_record_weight$Age..yr. < 75 ~ '52-75',
                                        TRUE ~ '75 + ')

# Fig 5
ggplot(data_record_weight, aes(x = Age_grp, y = Avg_Cr_Diff, color = as.factor(Flag))) +
  geom_boxplot() +
  scale_color_manual(values = c('#118ab2','#06d6a0', '#ffd166', '#ef476f'),
                     name = 'Weather Flag') +
  scale_y_continuous(name = 'Normalized Difference') +
  scale_x_discrete(name = 'Age Group') +
  facet_wrap(~Sex) +
```

```r
  theme_minimal()

# Normalized difference of weather parameters
# not used in the final report
data_record_weight <-data_record_weight %>%
  mutate(TdC_diff = (Td..C - mean_TdC_weighted)/mean_TdC_weighted,
         TwC_diff = (Tw..C - mean_TwC_weighted)/mean_TwC_weighted,
         TgC_diff = (Tg..C - mean_TgC_weighted)/mean_TgC_weighted,
         WBGT_diff = (WBGT - mean_WBGT_weighted)/mean_WBGT_weighted,
         SR_diff = (SR.W.m2 - mean_SR.W.m2_weighted)/mean_SR.W.m2_weighted,
         DP_diff = (DP - mean_DP_weighted)/mean_DP_weighted,
         Wind_diff = (Wind - mean_Wind_weighted)/mean_Wind_weighted)

# Correlation
cor_df <- data_record_weight %>%
  group_by(Age..yr. ) %>%
  summarise(Dry_T = cor(TdC_diff, Avg_Cr_Diff, use = 'complete.obs'),
            Wet_T = cor(TwC_diff, Avg_Cr_Diff, use = 'complete.obs'),
            Goble_T = cor(TgC_diff, Avg_Cr_Diff, use = 'complete.obs'),
            WBGT = cor(WBGT_diff, Avg_Cr_Diff, use = 'complete.obs'),
            SR = cor(SR_diff, Avg_Cr_Diff, use = 'complete.obs'),
            Dp = cor(DP_diff, Avg_Cr_Diff, use = 'complete.obs'),
            Wind = cor(Wind_diff, Avg_Cr_Diff, use = 'complete.obs'))

# convert to long format for Fig 6
cor_df_long <- pivot_longer(cor_df,
                            cols = c(Dry_T, Wet_T, Goble_T, WBGT, SR, Dp, Wind),
                            names_to = 'Type', values_to = 'Correlation')
cor_df_long$Type <- factor(cor_df_long$Type,
                           levels = c('Dry_T', 'Wet_T', 'Goble_T', 'WBGT',
                                      'SR', 'Dp', 'Wind'))

cor_df_long$Type <- case_when(cor_df_long$Type == 'Dry_T' ~ 'Dry Bulb Temperature',
                              cor_df_long$Type == 'Wet_T' ~ 'Wet Bulb Temperature',
                              cor_df_long$Type == 'Goble_T' ~ 'Globe Temperature',
                              cor_df_long$Type == 'WBGT' ~ 'WBGT',
                              cor_df_long$Type == 'SR' ~ 'Solar Radiation',
                              cor_df_long$Type == 'Dp' ~ 'Dew Point',
                              cor_df_long$Type == 'Wind' ~ 'Wind Speed')

# Fig6
ggplot(cor_df_long, aes(x = Age..yr., y = Correlation, color = as.factor(Type))) +
```

```r
  geom_point(size = 0.5) +
  geom_smooth(se = F) +
  scale_color_brewer(palette = 'Dark2', name = 'Weather Conditions') +
  scale_y_continuous(name = 'Correlation') +
  scale_x_continuous(name = 'Age (Year)') +
  theme_minimal()
#Age group
cor_df_long$Age_grp <- case_when(cor_df_long$Age..yr. < 26 ~ '< 26',
                                  cor_df_long$Age..yr. < 52 ~ '26-52',
                                  cor_df_long$Age..yr. < 75 ~ '52-75',
                                  TRUE ~ '75 + ')



# Fig 7
ggplot(cor_df_long, aes(x = Age_grp, y = Correlation, color = as.factor(Type))) +
  geom_boxplot() +
  scale_color_brewer(palette = 'Dark2', name = 'Weather Conditions') +
  scale_y_continuous(name = 'Correlation') +
  scale_x_discrete(name = 'Age Group') +
  theme_minimal()
# air quality data
air <- air[!is.na(air$aqi),]
air_avg <- air %>%
  group_by(date_local, marathon) %>%
  filter(sample_duration == "8-HR RUN AVG BEGIN HOUR") %>%
  summarise(mean_aqi = mean(aqi))

air_avg$Year <- as.numeric(substr(air_avg$date_local, 1, 4))
air_avg$Race <- case_when(
  air_avg$marathon == 'Boston' ~ 0,
  air_avg$marathon == 'Chicago' ~ 1,
  air_avg$marathon == 'NYC' ~ 2,
  air_avg$marathon == 'Twin Cities' ~ 3,
  air_avg$marathon == 'Grandmas' ~ 4
)

# merge air data with record data
data_record_weight_air <- left_join(data_record_weight, air_avg,
                                    by = c('Year', 'Race'))
# correlation between air and marathon performance
cor_air_df <- data_record_weight_air %>%
  group_by(Age..yr. ) %>%
```

```r
  summarise(Aqi_cor = cor(mean_aqi, Avg_Cr_Diff, use = 'complete.obs'))

cor_air_df$Age_grp <- case_when(cor_air_df$Age..yr. < 26 ~ '< 26',
                                 cor_air_df$Age..yr. < 52 ~ '26-52',
                                 cor_air_df$Age..yr. < 75 ~ '52-75',
                                 TRUE ~ '75 + ')

# fig 8
p1 <- ggplot(cor_air_df, aes(x = Age..yr., y = Aqi_cor)) +
  geom_point() +
  scale_y_continuous(name = 'Correlation') +
  scale_x_continuous(name = 'Age (Year)') +
  geom_smooth(se = F) +
  theme_minimal()

p2 <- ggplot(cor_air_df, aes(x = Age_grp, y = Aqi_cor)) +
  geom_boxplot() +
  scale_x_discrete(name = 'Age Group') +
  theme_minimal() +
  theme(axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
grid.arrange(p1, p2, ncol=2)
# linear model of weather parameters on marathon performance, condition on air quality
m_tdc <- lm(Avg_Cr_Diff ~ Td..C * mean_aqi, data = data_record_weight_air)
m_twc <- lm(Avg_Cr_Diff ~ Tw..C * mean_aqi, data = data_record_weight_air)
m_tgc <- lm(Avg_Cr_Diff ~ Tg..C * mean_aqi, data = data_record_weight_air)
m_WBGT <- lm(Avg_Cr_Diff ~ WBGT * mean_aqi, data = data_record_weight_air)
m_SR <- lm(Avg_Cr_Diff ~ SR.W.m2 * mean_aqi, data = data_record_weight_air)
m_DP <- lm(Avg_Cr_Diff ~ DP * mean_aqi, data = data_record_weight_air)
m_Wind <- lm(Avg_Cr_Diff ~ Wind * mean_aqi, data = data_record_weight_air)
m_rh <- lm(Avg_Cr_Diff ~ X.rh * mean_aqi, data = data_record_weight_air)

# select beta1 and beta3
coef_df <- round(summary(m_tdc)$coefficient[c(2,4),c(1,4)],3)
coef_df <- rbind(coef_df, round(summary(m_twc)$coefficient[c(2,4),c(1,4)],3))
coef_df <- rbind(coef_df, round(summary(m_tgc)$coefficient[c(2,4),c(1,4)],3))
coef_df <- rbind(coef_df, round(summary(m_WBGT)$coefficient[c(2,4),c(1,4)],3))
coef_df <- rbind(coef_df, round(summary(m_SR)$coefficient[c(2,4),c(1,4)],3))
coef_df <- rbind(coef_df, round(summary(m_DP)$coefficient[c(2,4),c(1,4)],3))
coef_df <- rbind(coef_df, round(summary(m_Wind)$coefficient[c(2,4),c(1,4)],3))
coef_df <- rbind(coef_df, round(summary(m_rh)$coefficient[c(2,4),c(1,4)],3))
```

```r
rownames(coef_df) <- c('Dry bulb temperature', 'Dry bulb temperature * Aqi',
                       'Wet bulb temperature ', 'Wet bulb temperature * Aqi',
                       'Black globe temperature', 'Black globe temperature * Aqi',
                       'Wet Bulb Globe Temperature', 'Wet Bulb Globe Temperature * Aqi',
                       'Solar Radiation', 'Solar Radiation * Aqi',
                       'Dew Point', 'Dew Point * Aqi',
                       'Wind Speed', 'Wind Speed * Aqi',
                       'Relative Humidity', 'Relative Humidity * Aqi')

#table3
coef_df <- as.data.frame(coef_df)
coef_df$Estimate <- ifelse(coef_df$Estimate ==0, '< 0.001', coef_df$Estimate)
coef_df$`Pr(>|t|)` <- ifelse(coef_df$`Pr(>|t|)` == 0, '< 0.001', coef_df$`Pr(>|t|)`)
coef_df$`Pr(>|t|)` <- ifelse(coef_df$`Pr(>|t|)` < 0.05,
                             paste0(coef_df$`Pr(>|t|)`, '*'), coef_df$`Pr(>|t|)`)
coef_df$Estimate <- as.character(coef_df$Estimate)
coef_df$`Pr(>|t|)` <- as.character(coef_df$`Pr(>|t|)`)
coef_df %>%
  kable(align = 'ccc',col.names = c('Parameter', 'Estimate', 'P-value')) %>%
  kable_styling()
```