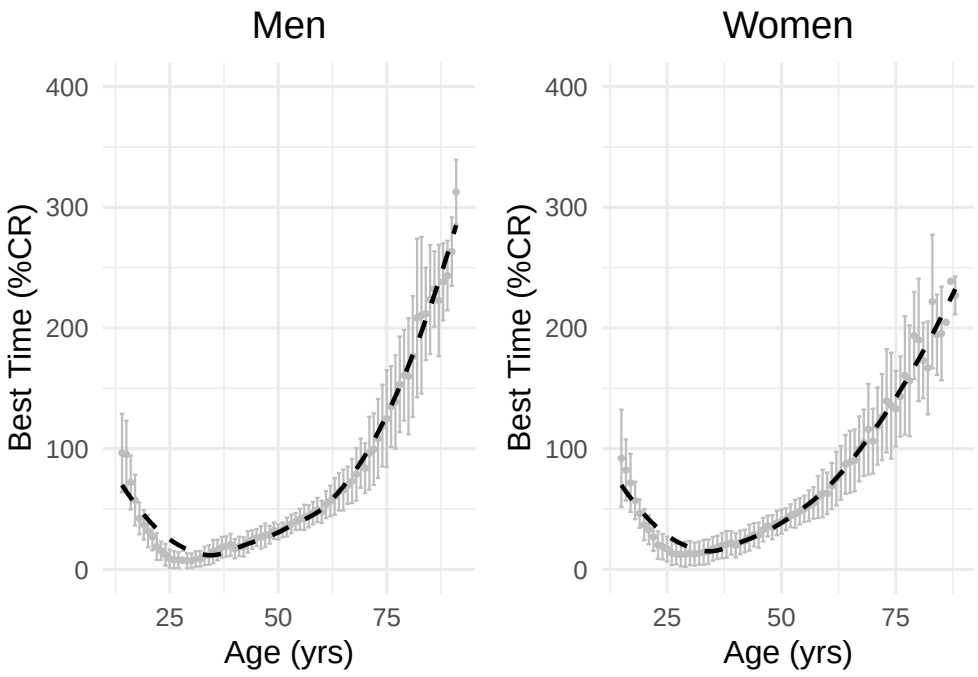
Project1 Codebook

Yingxi Kong

2024-10-01

Table 1: Summary of Marathon Performance by Age Group and Sex

Age Group	Sex	N	Min Performance	Mean Performance	Median Performance	Max Performance
Younger	Female	500	2.039	51.777	47.583	211.095
Younger	Male	547	-1.074	46.954	39.385	159.535
Lower Age	Female	960	-1.816	15.933	14.164	53.263
Lower Age	Male	959	-2.251	10.112	7.518	62.806
Lower-Mid Age	Female	1536	-1.419	22.329	22.548	60.830
Lower-Mid Age	Male	1536	-0.499	18.943	19.010	89.271
Upper-Mid Age	Female	1435	8.873	52.005	49.190	157.858
Upper-Mid Age	Male	1439	7.880	41.257	39.577	96.567
Highest Age	Female	1021	36.187	113.259	104.046	336.347
Highest Age	Male	1631	27.918	107.439	90.900	419.958



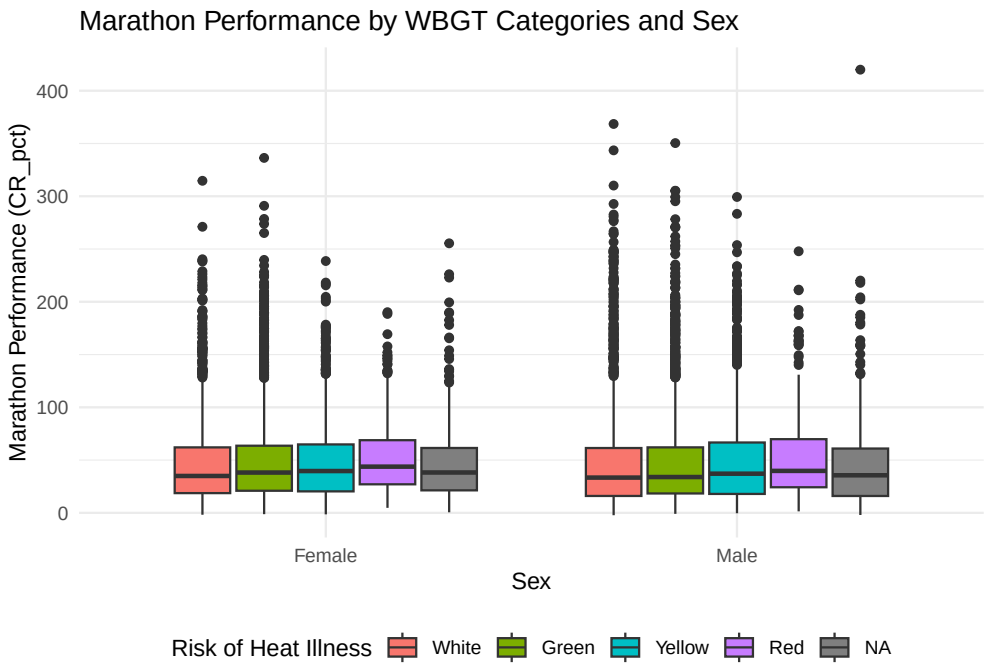
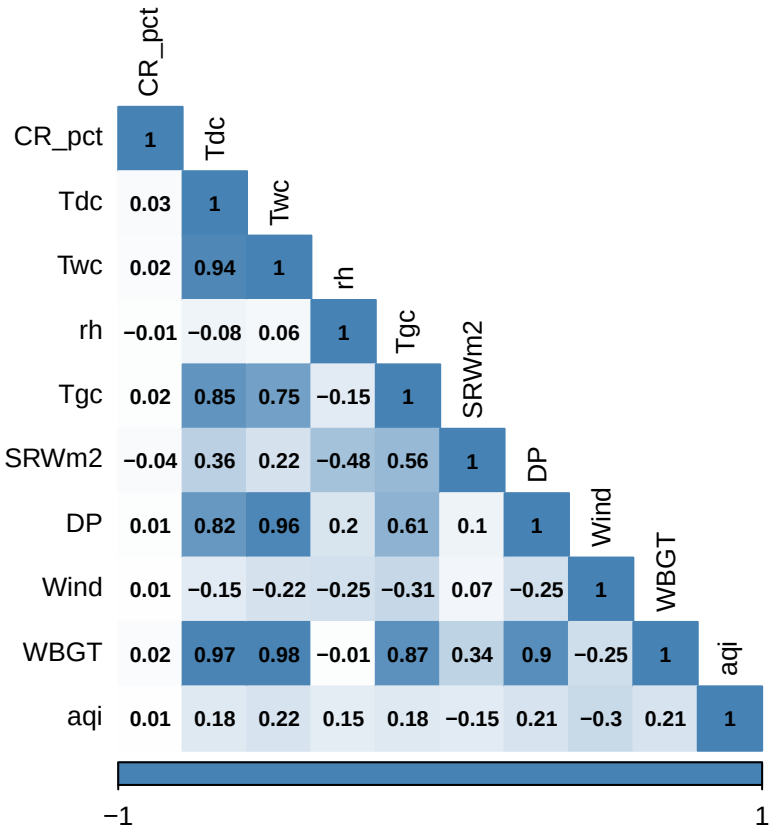
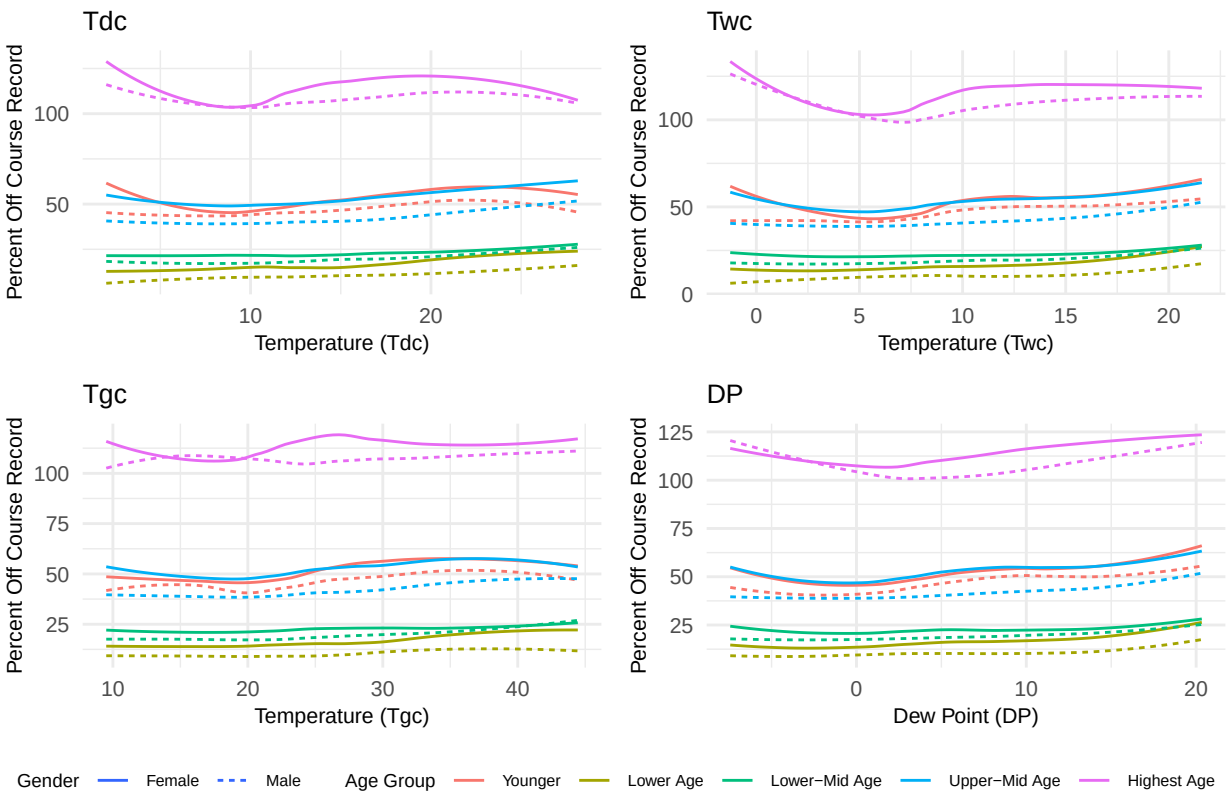

Marathon Performance by WBGT Categories and Sex
Marathon Performance (CR_pct)
400
300
200
100
0
Female
Male
Sex
Risk of Heat Illness
White
Green
Yellow
Red
NA

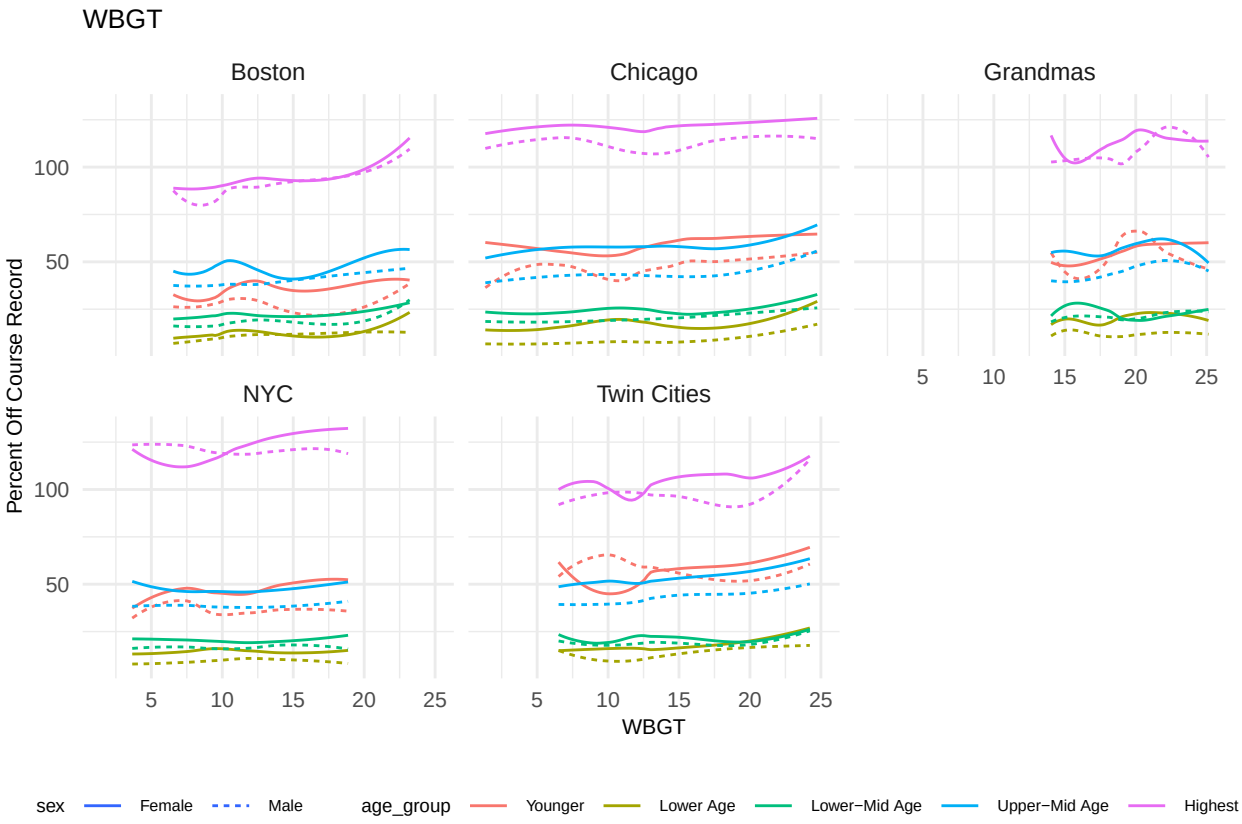
Table 1: Summary Statistics by Race

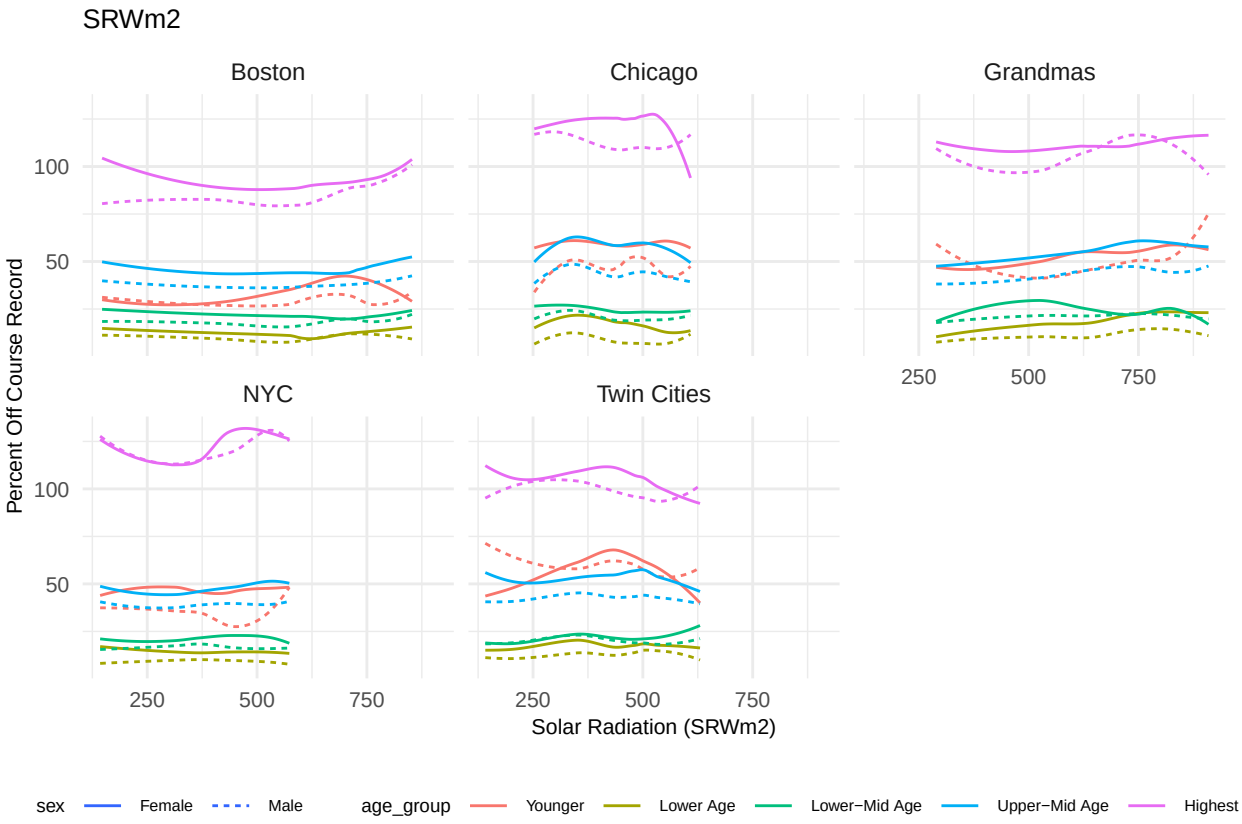
Characteristic	Race				
	Boston Marathon, N = 2,088[1]	**Chicago Marathon**, N = 2,553[1]	**Grandma's Marathon**, N = 2,000[1]	**New York City Marathon**, N = 2,930[1]	**Twin Cities Marathon**, N = 1,993[1]
Gender					
Female	984 (47%)	1,210 (47%)	934 (47%)	1,402 (48%)	922 (46%)
Male	1,104 (53%)	1,343 (53%)	1,066 (53%)	1,528 (52%)	1,071 (54%)
flag					
Missing	0 (0%)	126 (4.9%)	116 (5.8%)	131 (4.5%)	118 (5.9%)
WBGT < 10C	1,040 (50%)	732 (29%)	0 (0%)	1,394 (48%)	587 (29%)
WBGT > 18-23C	115 (5.5%)	120 (4.7%)	945 (47%)	504 (17%)	338 (17%)
WBGT > 23-28C	123 (5.9%)	116 (4.5%)	237 (12%)	0 (0%)	116 (5.8%)
WBGT 10-18C	810 (39%)	1,459 (57%)	702 (35%)	901 (31%)	834 (42%)
Age	47 (17)	46 (18)	44 (18)	50 (19)	45 (17)
Percent off current course record	41 (34)	52 (47)	48 (40)	55 (56)	46 (37)
Dry bulb temperature	11.6 (5.9)	12.4 (6.1)	18.9 (3.3)	11.7 (4.7)	13.1 (5.5)
Missing	0	126	116	131	118
Wet bulb temperature	7.6 (3.8)	8.5 (5.7)	14.9 (2.5)	7.6 (5.0)	9.9 (5.4)
Missing	0	126	116	131	118
Percent relative humidity	36 (34)	60 (10)	49 (34)	27 (30)	42 (34)
Missing	0	126	116	131	118
Black globe temperature	24 (8)	25 (6)	32 (8)	21 (6)	25 (7)
Missing	0	126	116	131	118
Solar radiation in Watts	650 (187)	460 (95)	677 (191)	401 (131)	436 (139)
Missing	0	126	116	131	118
Dew Point	3 (4)	5 (7)	12 (3)	3 (7)	6 (7)
Missing	0	126	116	131	118
Wind	12.0 (4.5)	8.2 (3.2)	9.2 (2.9)	11.2 (4.6)	8.8 (3.2)
Missing	0	126	116	131	118
WBGT	11.3 (4.5)	12.1 (5.8)	18.6 (3.2)	10.7 (4.9)	13.2 (5.4)
Missing	0	126	116	131	118
Air Quality Index	3.8 (2.3)	4.5 (3.4)	2.9 (1.9)	4.0 (5.3)	5.8 (3.5)

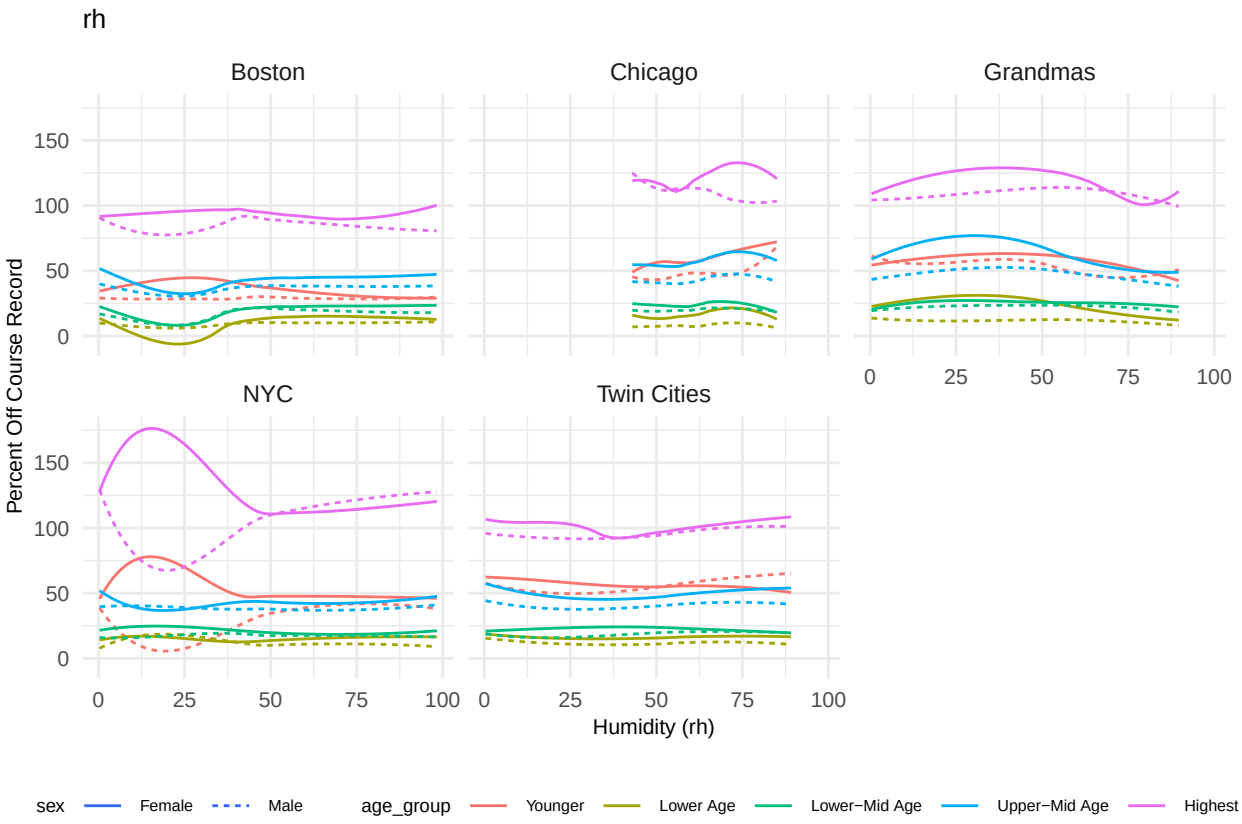
[1]Mean (SD) for continuous; n (%) for categorical

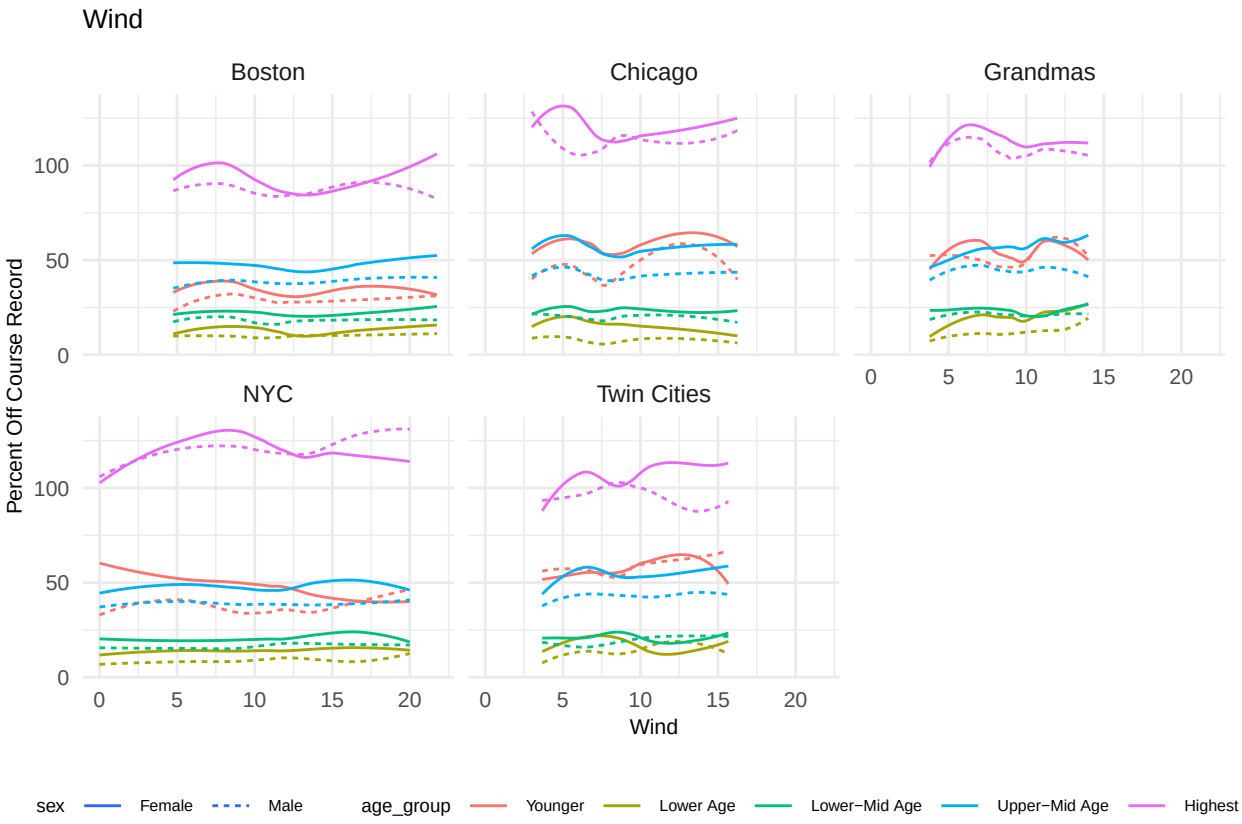

CR_pct
Tdc
Twc
rh
Tgc
SRWm2
DP
Wind
WBGT
aqi
1
0.03 1
0.02 0.94 1
−0.01 −0.08 0.06 1
0.02 0.85 0.75 −0.15 1
−0.04 0.36 0.22 −0.48 0.56 1
0.01 0.82 0.96 0.2 0.61 0.1 1
0.01 −0.15 −0.22 −0.25 −0.31 0.07 −0.25 1
0.02 0.97 0.98 −0.01 0.87 0.34 0.9 −0.25 1
0.01 0.18 0.22 0.15 0.18 −0.15 0.21 −0.3 0.21 1
−1
1

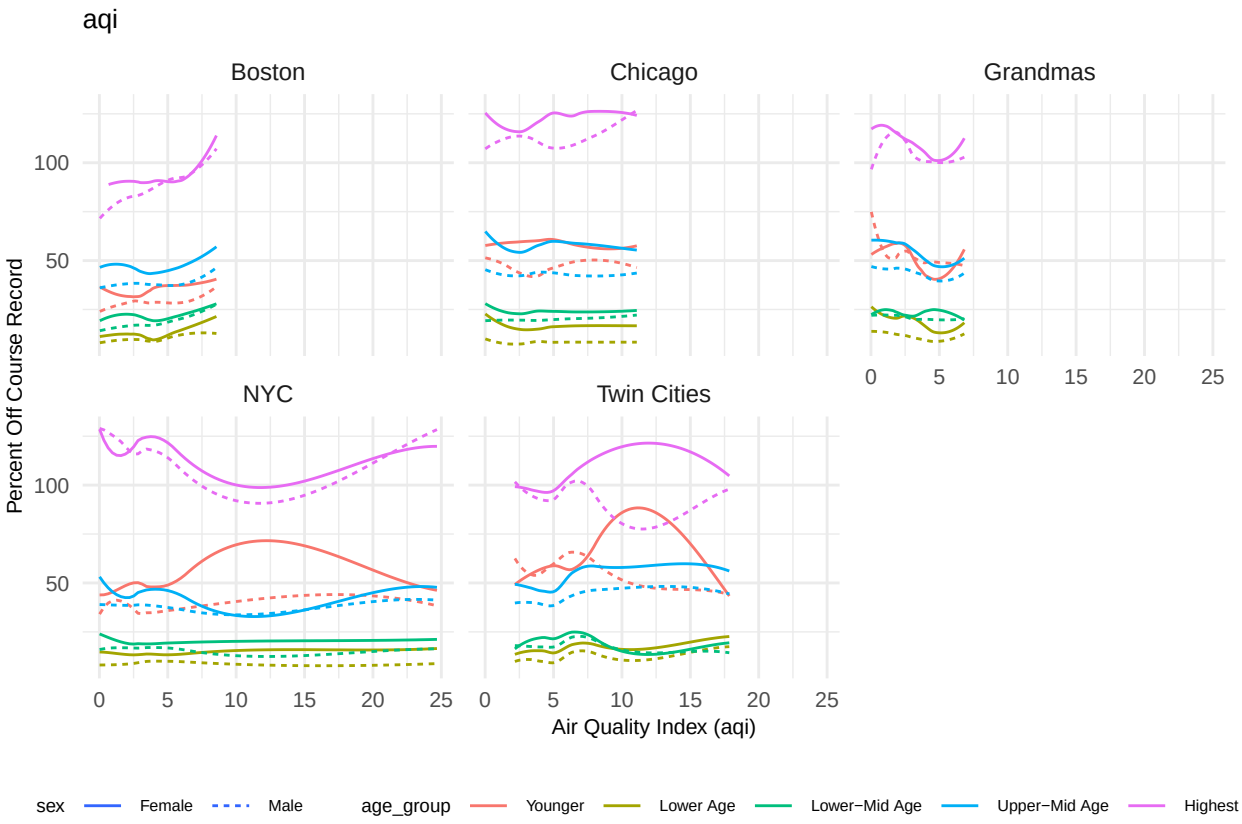

Overall Performance vs. Temperature Conditions
Tdc
Twc
Tgc
DP
Percent Off Course Record
Temperature (Tdc)
Temperature (Twc)
Temperature (Tgc)
Dew Point (DP)
Gender
Female
Male
Age Group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age

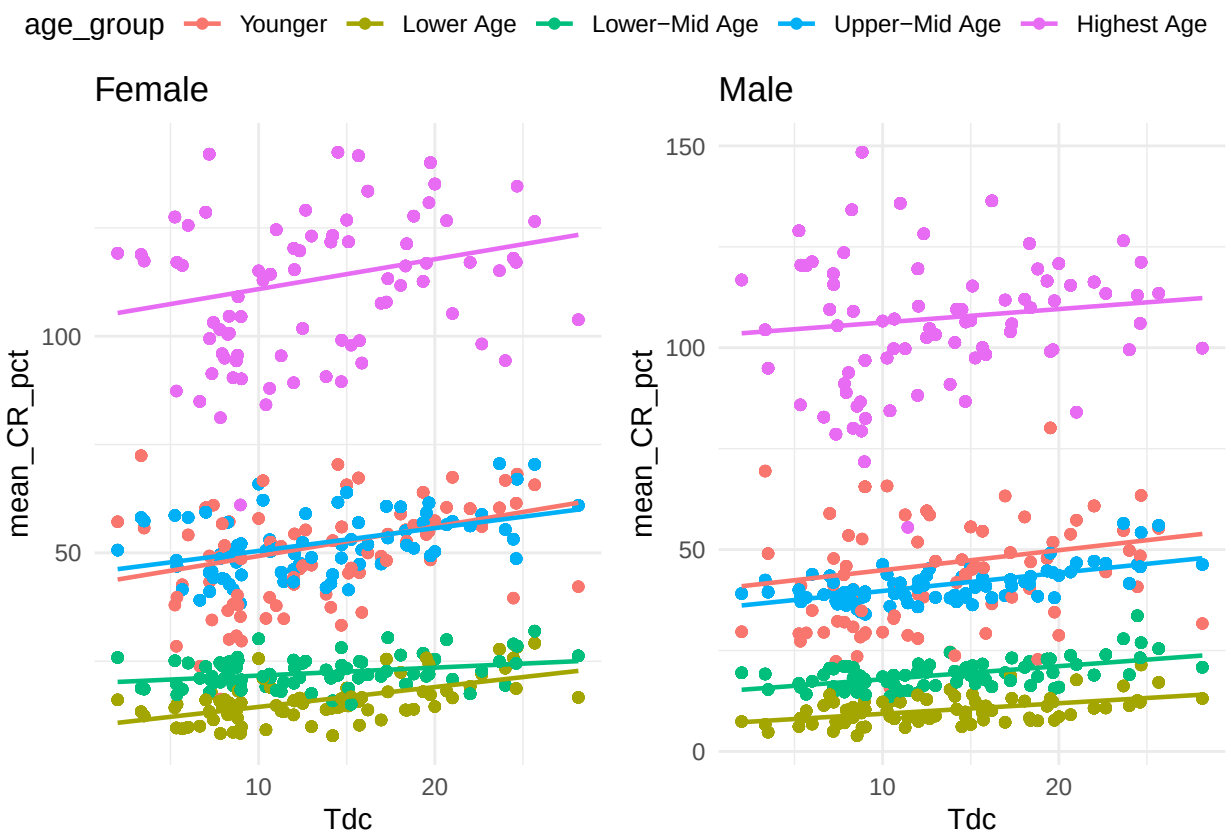

WBGT
Boston
Chicago
Grandmas
NYC
Twin Cities
Percent Off Course Record
100
50
5
10
15
20
25
WBGT
sex
Female
Male
age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest

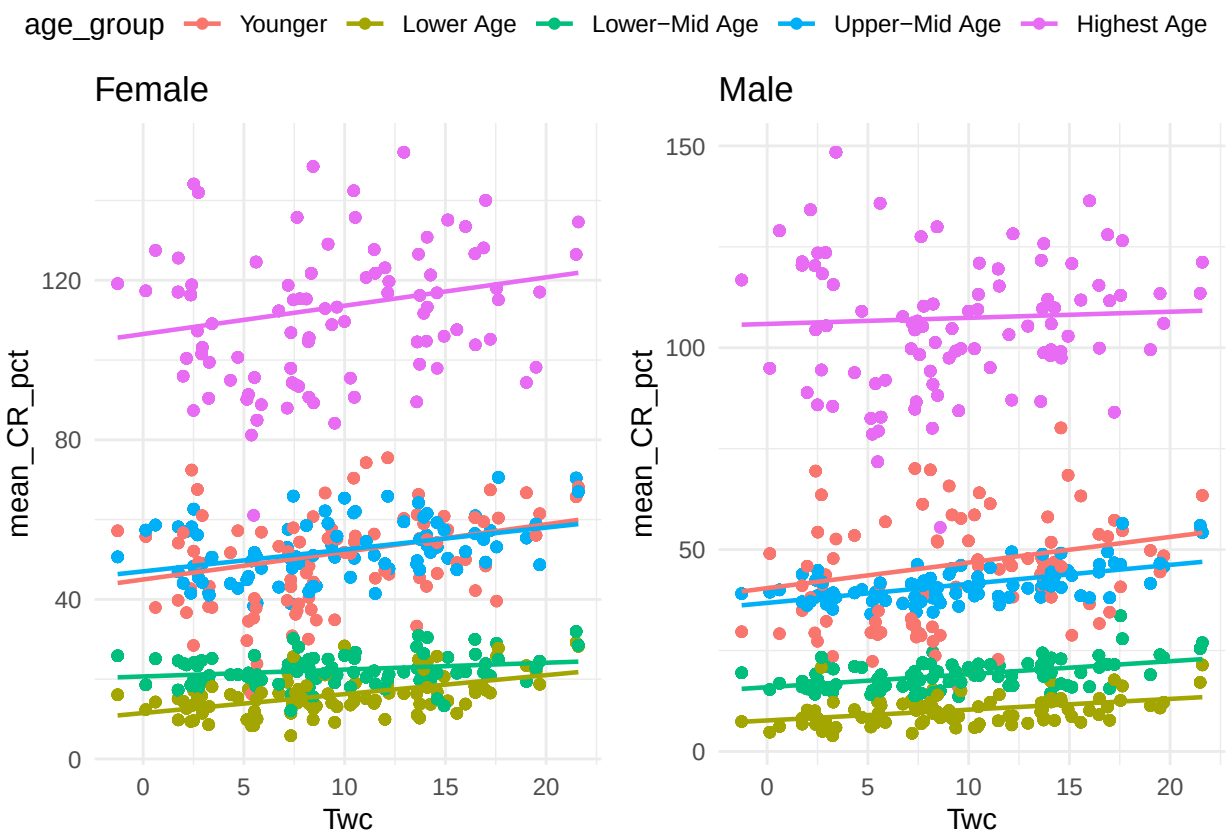

SRWm2
Boston
Chicago
Grandmas
NYC
Twin Cities
Percent Off Course Record
100
50
250
500
750
Solar Radiation (SRWm2)
sex
Female
Male
age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest

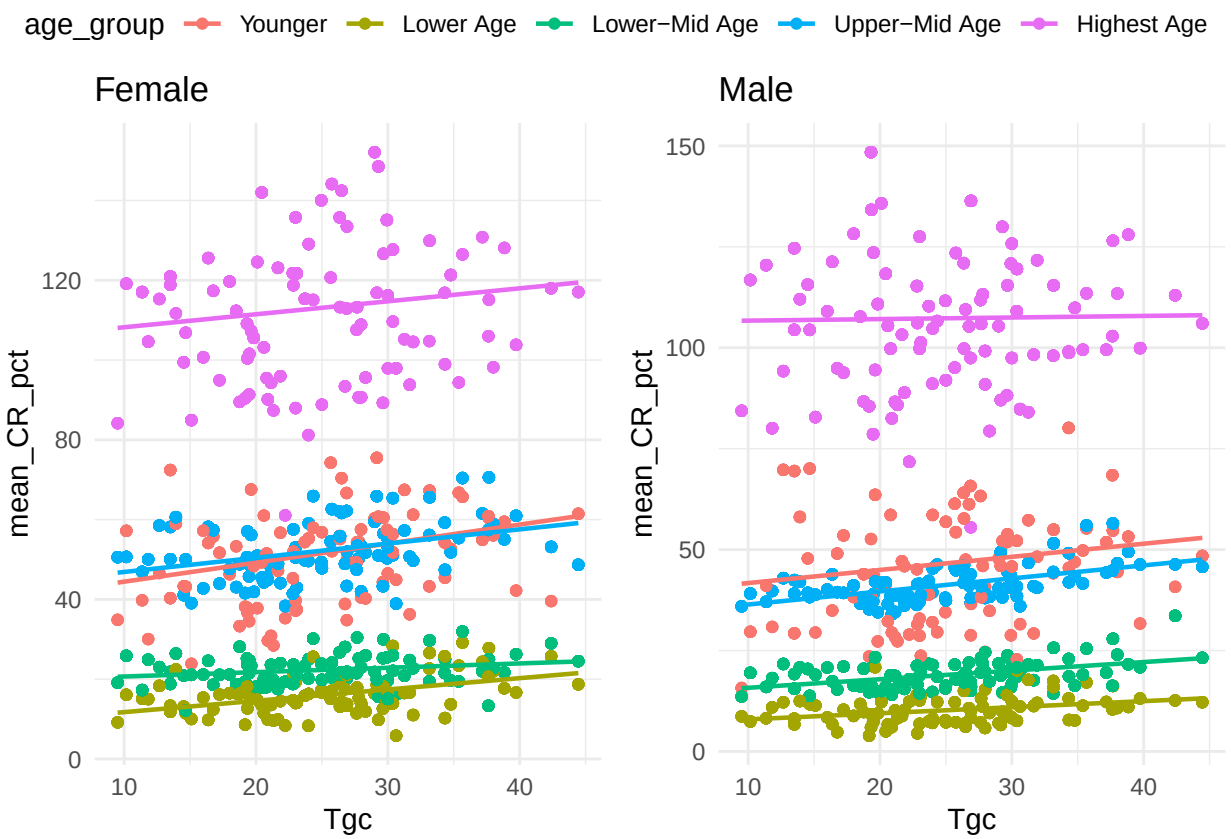

rh
Boston
Chicago
Grandmas
NYC
Twin Cities
Percent Off Course Record
Humidity (rh)
0
50
100
150
0
25
50
75
100
sex
Female
Male
age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest

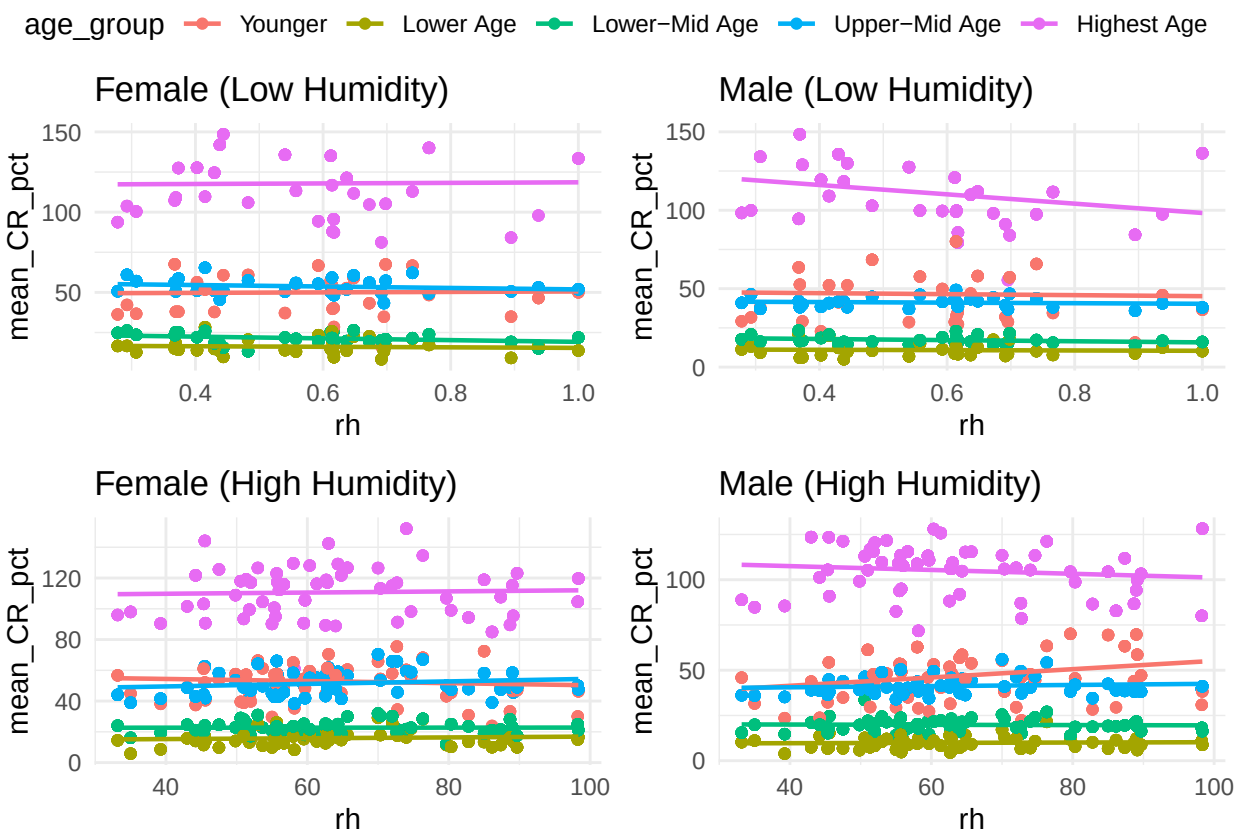

Wind
Boston
Chicago
Grandmas
NYC
Twin Cities
Percent Off Course Record
100
50
0
0
5
10
15
20
Wind
sex
Female
Male
age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest

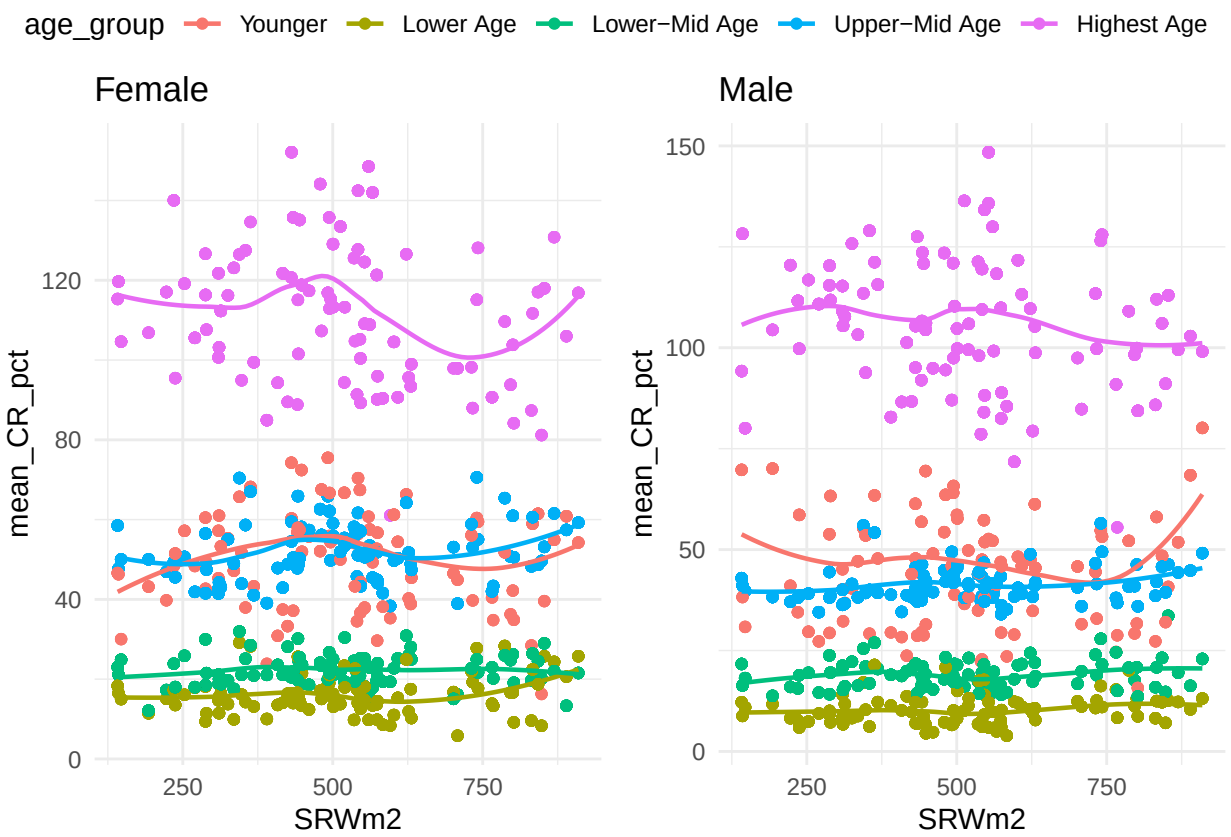

aqi
Boston
Chicago
Grandmas
NYC
Twin Cities
Percent Off Course Record
100
50
0
5
10
15
20
25
Air Quality Index (aqi)
sex
Female
Male
age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest

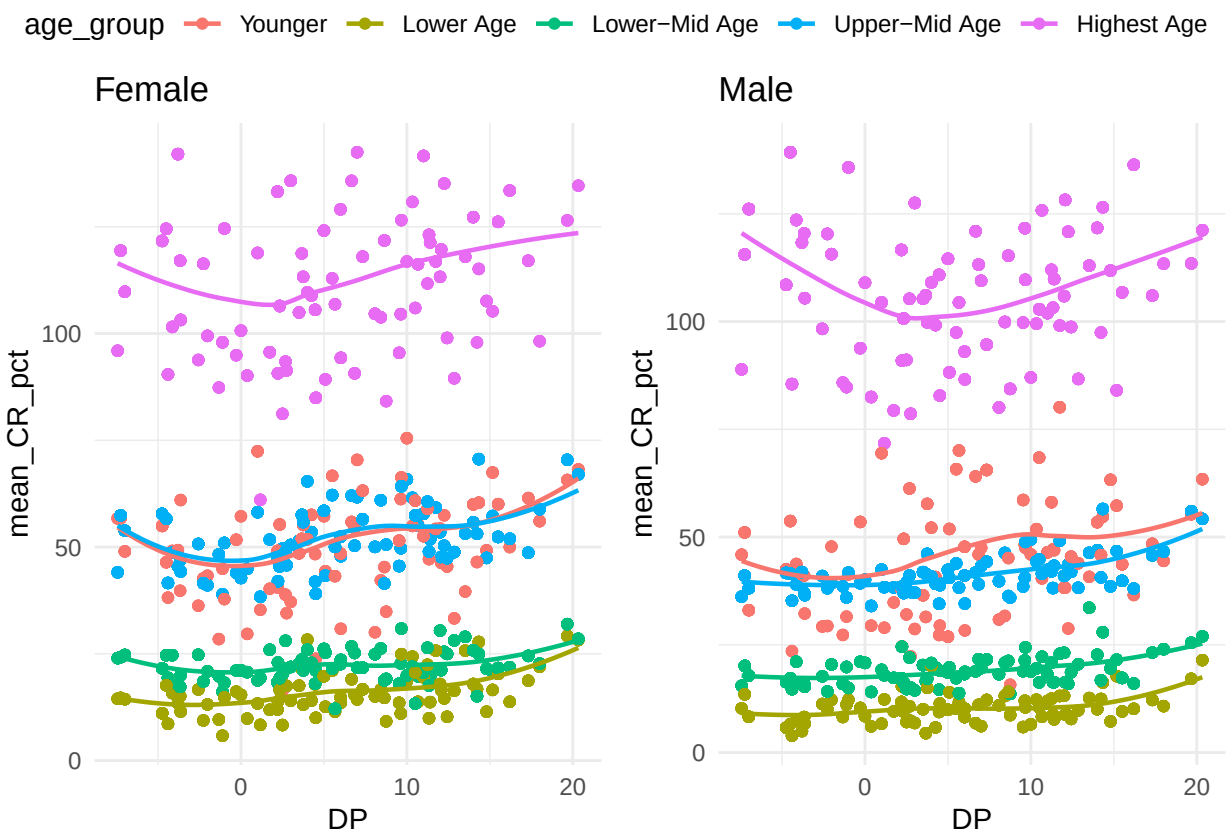

age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
Tdc
150
100
50
0
10
20

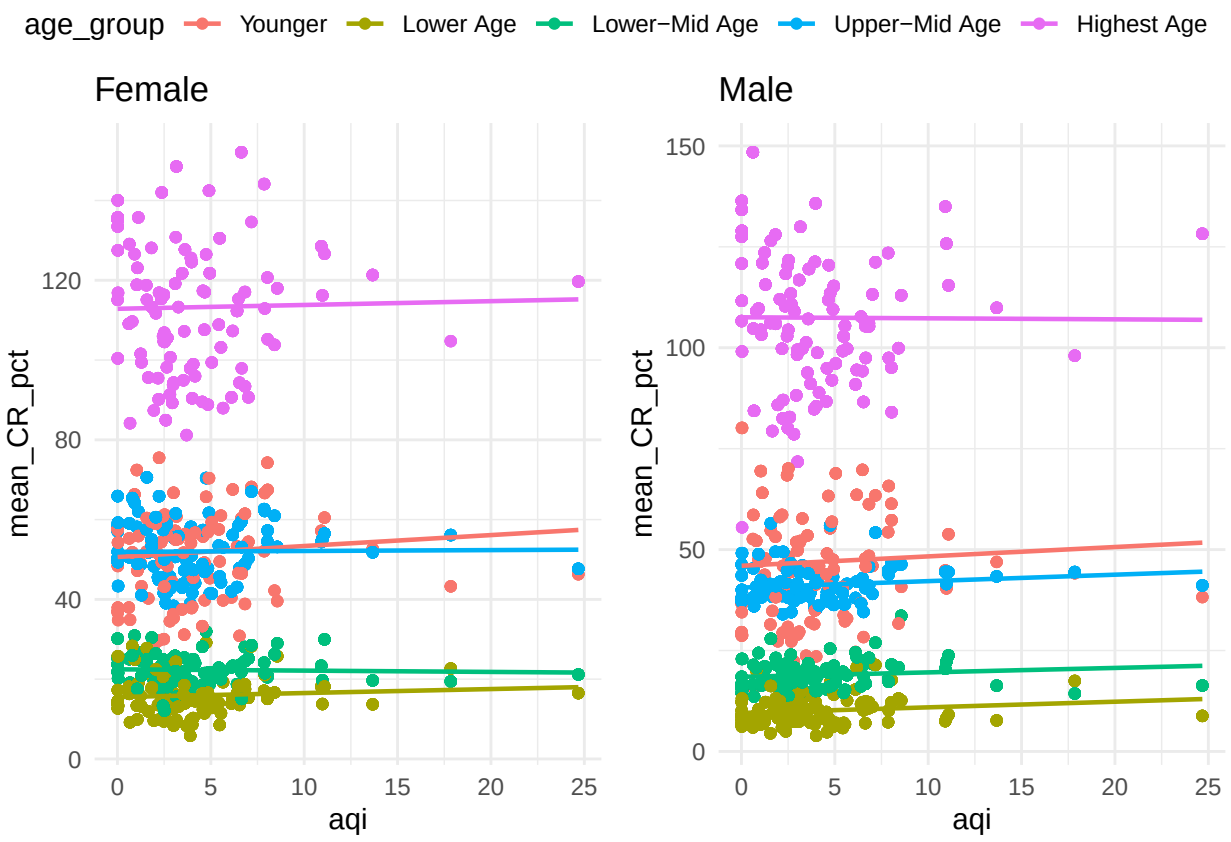

age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
Twc
0
40
80
120
50
100
150
0
5
10
15
20

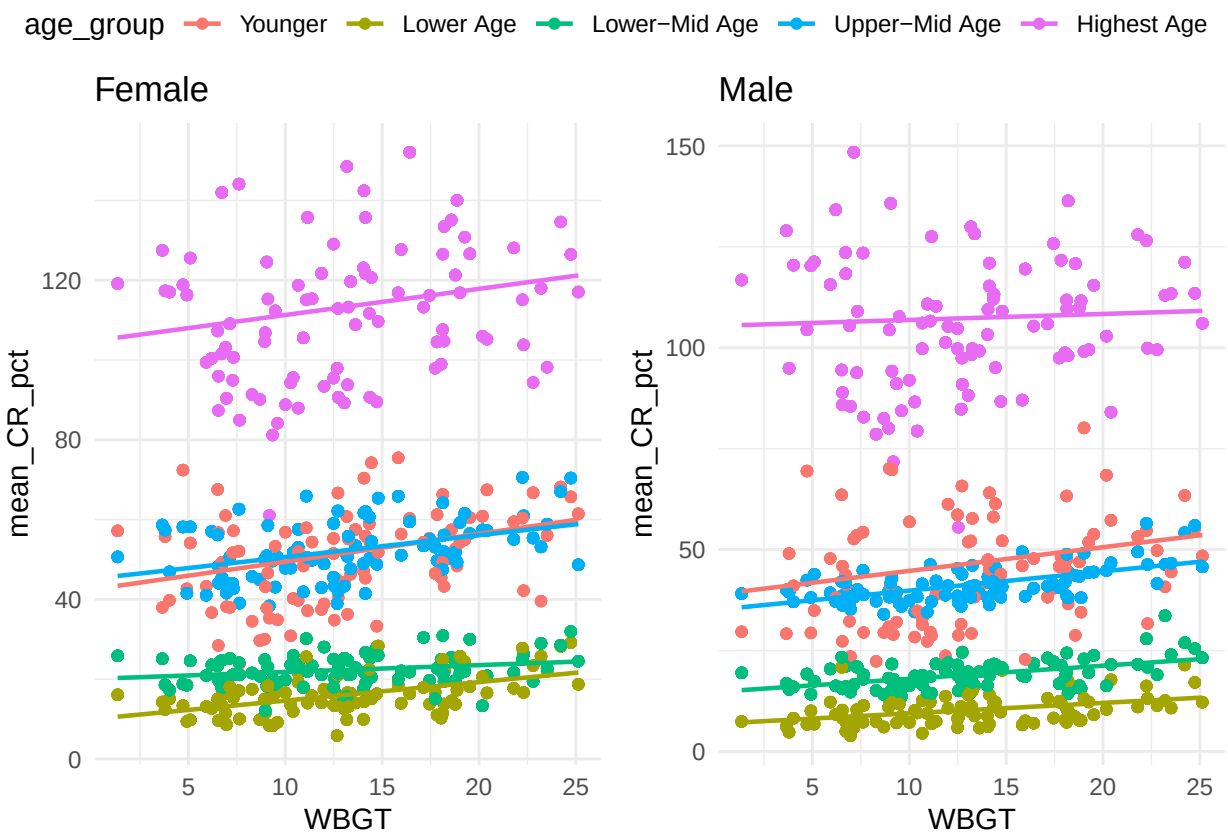

age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
Tgc
0
40
80
120
50
100
150
10
20
30
40

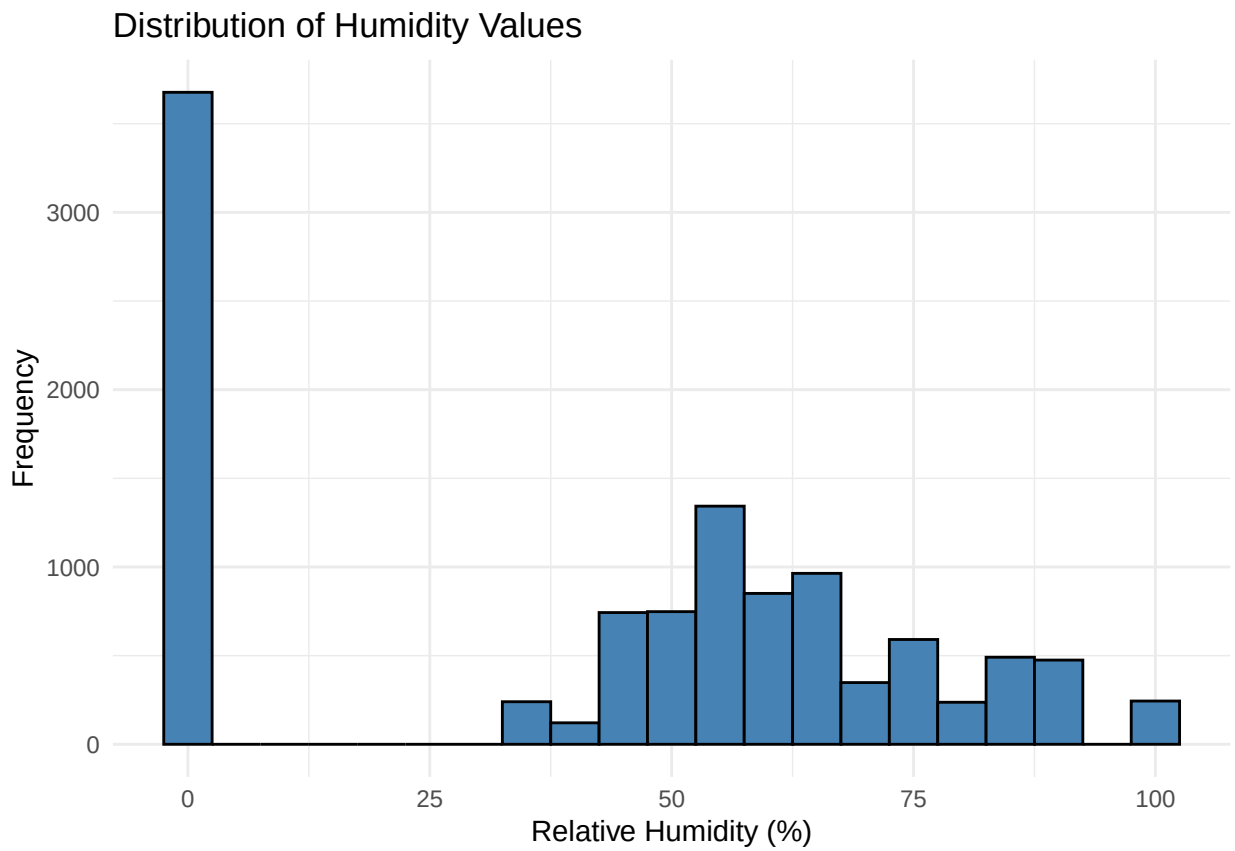

age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female (Low Humidity)
Male (Low Humidity)
Female (High Humidity)
Male (High Humidity)
mean_CR_pct
rh


age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
SRWm2
0
40
80
120
50
100
150
250
500
750

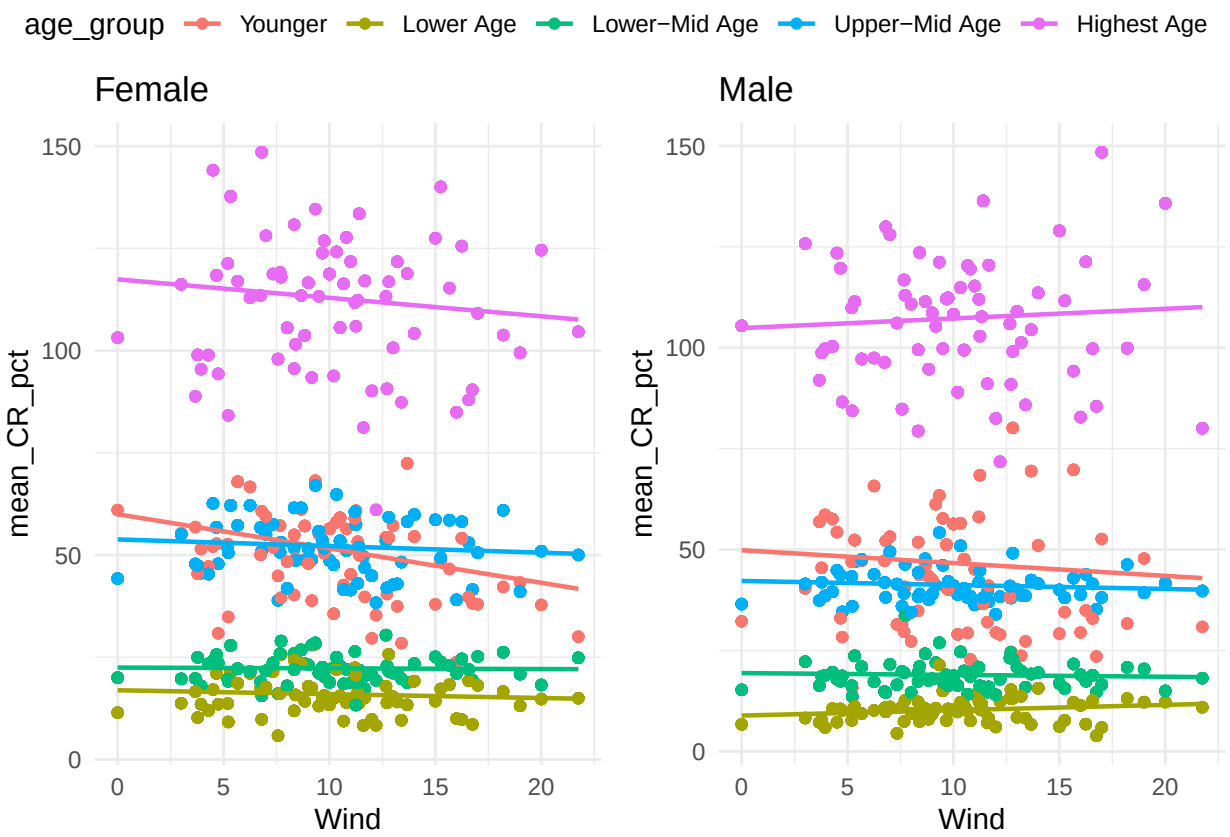

age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
DP
0
50
100
0
10
20


age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
aqi
0
40
80
120
50
100
150
0
5
10
15
20
25


age_group
Younger
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Female
Male
mean_CR_pct
WBGT
0
40
80
120
50
100
150
5
10
15
20
25




```
## data$sex: Female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   30.00   45.00   44.99   59.00   88.00
## ------------------------------------------------------------
## data$sex: Male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   32.00   48.00   47.81   64.00   91.00
```




```
##
## Call:
## lm(formula = CR_pct ~ Tdc * age + Tdc * sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.016 -19.486  -9.797   7.496 259.571
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43.950367   2.195273 -20.020  < 2e-16 ***
## Tdc           1.012743   0.150244   6.741 1.66e-11 ***
## age           1.915486   0.042287  45.297  < 2e-16 ***
## sexMale      -4.441930   1.515898  -2.930  0.00339 **
## Tdc:age      -0.011736   0.002924  -4.013 6.02e-05 ***
## Tdc:sexMale  -0.026900   0.104369  -0.258  0.79662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.71 on 11067 degrees of freedom
##   (491 observations deleted due to missingness)
## Multiple R-squared:  0.4977, Adjusted R-squared:  0.4975
## F-statistic:  2193 on 5 and 11067 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = CR_pct ~ Tdc * age + Tdc * sex + Twc * age + Twc *
```

```
##     sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.786 -19.519  -9.780   7.571 256.137
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.406082   2.455643 -15.640  < 2e-16 ***
## Tdc          -1.115830   0.456556  -2.444  0.01454 *
## age           1.791633   0.047543  37.684  < 2e-16 ***
## sexMale      -4.481959   1.691066  -2.650  0.00805 **
## Twc           2.423433   0.491634   4.929 8.37e-07 ***
## Tdc:age       0.035342   0.008818   4.008 6.16e-05 ***
## Tdc:sexMale  -0.008124   0.315035  -0.026  0.97943
## age:Twc      -0.053581   0.009470  -5.658 1.57e-08 ***
## sexMale:Twc  -0.022967   0.340276  -0.067  0.94619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.66 on 11064 degrees of freedom
##   (491 observations deleted due to missingness)
## Multiple R-squared:  0.4992, Adjusted R-squared:  0.4988
## F-statistic:  1378 on 8 and 11064 DF,  p-value: < 2.2e-16
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: CR_pct ~ sex + rh + age + SRWm2 + Wind + WBGT + (1 | age)
##    Data: data
##
## REML criterion at convergence: 96017.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.8872 -0.4724 -0.0250  0.3994  9.6217
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  age      (Intercept) 1953.6   44.20
##  Residual              325.8   18.05
## Number of obs: 11073, groups:  age, 78
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -5.875e+01  1.277e+01  7.547e+01  -4.600 1.67e-05 ***
## sexMale     -9.901e+00  3.472e-01  1.099e+04 -28.513  < 2e-16 ***
## rh          -1.212e-02  6.379e-03  1.099e+04  -1.900  0.05748 .
## age          2.668e+00  2.234e-01  7.529e+01  11.939  < 2e-16 ***
## SRWm2       -3.085e-03  1.135e-03  1.099e+04  -2.719  0.00655 **
## Wind        -3.952e-02  4.516e-02  1.099e+04  -0.875  0.38146
## WBGT         4.755e-01  3.444e-02  1.099e+04  13.807  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##         (Intr) sexMal rh     age    SRWm2  Wind
## sexMale -0.010
## rh      -0.051 -0.003
## age     -0.917 -0.006  0.007
## SRWm2   -0.046 -0.012  0.491  0.006
## Wind    -0.044  0.002  0.202 -0.003 -0.044
## WBGT    -0.025 -0.001 -0.121  0.001 -0.378  0.257
```

Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
library(tidyverse)
library(ggplot2)
library(visdat)
library(gtsummary)
library(kableExtra)
library(ggpubr)
library(gt)
library(car)
library(lme4)
library(lmerTest)
library(corrplot)

data <- read.csv("project1.csv")
course_record <- read.csv("course_record.csv")
aqi_values <- read.csv("aqi_values.csv")
marathon_dates <-  read.csv("marathon_dates.csv")

colnames(data) <- c("race", "year", "sex", "flag", "age", "CR_pct", "Tdc", "Twc", "rh", "Tgc", "SRWm2",

data$flag <- case_when(data$flag == "" ~ NA, TRUE ~ data$flag)
data$flag <- as.factor(data$flag)

course_record$Race <- case_when(course_record$Race == "B" ~ 0,
                                course_record$Race == "C" ~ 1,
                                course_record$Race == "NY" ~ 2,
                                course_record$Race == "TC" ~ 3,
                                course_record$Race == "D" ~ 4,
                                TRUE ~ NA)
course_record$Gender <- case_when(course_record$Gender == "M" ~ 1,
                                  course_record$Gender == "F" ~ 0,
                                  TRUE ~ NA)

data <- data %>%
  left_join(course_record, join_by("sex" == "Gender", "race" == "Race", "year" == "Year"))

data$sex <- as.factor(data$sex)
data$race <- case_when(data$race == 0 ~ "Boston",
                       data$race == 1 ~ "Chicago",
                       data$race == 2 ~ "NYC",
                       data$race == 3 ~ "Twin Cities",
                       data$race == 4 ~ "Grandmas")
```

```r
data <- data %>% left_join(marathon_dates, by = c("race" = "marathon", "year" = "year"))

aqi_values_sum <- aqi_values %>%
  group_by(marathon, date_local) %>%
  summarize(ave_arithmetic_mean = mean(arithmetic_mean, na.rm = TRUE), .groups = "drop")

data <- data %>%
  mutate(date = as.Date(date))
aqi_values_sum <- aqi_values_sum %>%
  mutate(date_local = as.Date(date_local))

# Joining with the main dataset
data <- data %>%
  left_join(aqi_values_sum, by = c("race" = "marathon", "date" = "date_local"))

colnames(data)[17] <- "aqi"

# Missing Pattern
# vis_dat(data)

# Classify each observation to age groups by gender's quantile value
# data <- data %>%
#   group_by(sex, .groups = "drop") %>%
#   mutate(age_group = cut(age, breaks = quantile(age, probs = seq(0, 1, 0.25), na.rm = TRUE),
#                                  include.lowest = TRUE,
#                                  labels = c("Lower Age", "Lower-Mid Age",
#                                             "Upper-Mid Age", "Highest Age")))


data <- data %>%
  group_by(sex) %>%
  mutate(age_group = cut(age,
                         breaks = c(0, 21, 31, 47, 62, Inf),
                         include.lowest = TRUE,
                         labels = c("Younger", "Lower Age", "Lower-Mid Age",
                                    "Upper-Mid Age", "Highest Age"))) %>%
  ungroup()

# Summary Table by Age group and Sex
summary_table <- data %>%
  group_by(age_group, sex) %>%
  summarize(N = n(),
            min_performance = round(min(CR_pct, na.rm = TRUE), 3),
            mean_performance = round(mean(CR_pct, na.rm = TRUE), 3),
            median_performance = round(median(CR_pct, na.rm = TRUE), 3),
            max_performance = round(max(CR_pct, na.rm = TRUE), 3))
summary_table$sex <- ifelse(summary_table$sex == 1, "Male", "Female")

knitr::kable(summary_table,
             col.names = c("Age Group", "Sex", "N", "Min Performance",
                           "Mean Performance", "Median Performance", "Max Performance"),
             caption = "Summary of Marathon Performance by Age Group and Sex") %>%
  kable_styling(latex_options = "HOLD_position",
```

```
                font_size = 8)

# male summary
male_summary <- data %>%
  filter(sex == 1) %>%
  group_by(age) %>%
  summarise(mean_CR = mean(CR_pct, na.rm = TRUE),
            se_CR = sd(CR_pct, na.rm = TRUE))

# create the plot
ageplot_male <- ggplot(male_summary, aes(x = age, y = mean_CR)) +
  geom_point(color = "grey", size = 1) +
  geom_errorbar(aes(ymin = mean_CR - se_CR, ymax = mean_CR + se_CR), width = 1, color = "grey") +
  geom_smooth(se = FALSE, color = "black", size = 1, method = "loess", linetype = 2) +
  labs(title = "Men", x = "Age (yrs)", y = "Best Time (%CR)") +
  ylim(0, 400) +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5))

# women summary
women_summary <- data %>%
  filter(sex == 0) %>%
  group_by(age) %>%
  summarise(mean_CR = mean(CR_pct, na.rm = TRUE),
            se_CR = sd(CR_pct, na.rm = TRUE))

# create the plot
ageplot_female <- ggplot(women_summary, aes(x = age, y = mean_CR)) +
  geom_point(color = "grey", size = 1) +
  geom_errorbar(aes(ymin = mean_CR - se_CR, ymax = mean_CR + se_CR), width = 1, color = "grey") +
  geom_smooth(se = FALSE, color = "black", size = 1, method = "loess", linetype = 2) +
  labs(title = "Women", x = "Age (yrs)", y = "Best Time (%CR)") +
  ylim(0, 400) +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5))

# merge the two plots together
ggarrange(ageplot_male, ageplot_female)
data$flag <- factor(data$flag, levels = c("White", "Green", "Yellow", "Red", "Black", NA))
data$sex <- ifelse(data$sex == 0, "Female", "Male")
ggplot(data, aes(x = sex, y = CR_pct, fill = flag)) +
  geom_boxplot() +
  ggtitle("Marathon Performance by WBGT Categories and Sex") +
  theme_minimal() +
  labs(x = "Sex", y = "Marathon Performance (CR_pct)", fill = "Risk of Heat Illness") +
  theme(legend.position = "bottom")
tbl_summary <- data %>%
  select(-c("year", "CR", "age_group", "date")) %>%
  mutate(
    race = case_when(race == "Boston" ~ "Boston Marathon",
                     race == "Chicago" ~ "Chicago Marathon",
                     race == "NYC" ~ "New York City Marathon",
                     race == "Twin Cities" ~ "Twin Cities Marathon",
```

```
                      race == "Grandmas" ~ "Grandma's Marathon",
                      TRUE ~ "Missing"),
    flag = case_when(flag == 'White' ~ "WBGT < 10C",
                      flag == 'Green' ~ "WBGT 10-18C",
                      flag == 'Yellow' ~ "WBGT > 18-23C",
                      flag == 'Red' ~ "WBGT > 23-28C",
                      TRUE ~ "Missing")) %>%
  tbl_summary(by = race,
              label = list(age ~ "Age",
                           sex ~ "Gender",
                           CR_pct ~ "Percent off current course record ",
                           Tdc ~ "Dry bulb temperature",
                           Twc ~ "Wet bulb temperature",
                           rh ~ "Percent relative humidity",
                           Tgc ~ "Black globe temperature",
                           SRWm2 ~ "Solar radiation in Watts",
                           DP ~ "Dew Point",
                           aqi ~ "Air Quality Index"),
              statistic = all_continuous() ~ "{mean} ({sd})",
              missing = "ifany",
              missing_text = "Missing") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Race**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
  bold_labels()

tbl_summary %>%
  as_gt() %>%
  cols_align(align = "center", columns = everything()) %>%
  tab_options(
    table.width = pct(100),
    table.align = "center"
  ) %>%
  cols_width(
    everything() ~ px(75)
  ) %>%
  tab_header(
    title = md("Table 1: Summary Statistics by Race")
  ) %>%
  tab_style(
    style = cell_text(size = px(8)),
    locations = cells_title(groups = "title")
  ) %>%
  tab_style(
    style = cell_text(size = px(7)),
    locations = cells_body()
  ) %>%
  tab_style(
    style = cell_text(size = px(7)),
    locations = cells_column_labels()
  )
# tbl_summary_female <- data %>%
#   filter(sex == "Female") %>%
#   select(-c(year, sex, CR, age_group, date)) %>%
```

```
#   mutate(
#     race = case_when(race == "Boston" ~ "Boston Marathon",
#                      race == "Chicago" ~ "Chicago Marathon",
#                      race == "NYC" ~ "New York City Marathon",
#                      race == "Twin Cities" ~ "Twin Cities Marathon",
#                      race == "Grandmas" ~ "Grandma's Marathon",
#                      TRUE ~ "Missing"),
#     flag = case_when(flag == 'White' ~ "WBGT < 10C",
#                      flag == 'Green' ~ "WBGT 10-18C",
#                      flag == 'Yellow' ~ "WBGT > 18-23C",
#                      flag == 'Red' ~ "WBGT > 23-28C",
#                      TRUE ~ "Missing")) %>%
#   tbl_summary(by = race,
#               label = list(age ~ "Age",
#                            CR_pct ~ "Percent off current course record ",
#                            Tdc ~ "Dry bulb temperature",
#                            Twc ~ "Wet bulb temperature",
#                            rh ~ "Percent relative humidity",
#                            Tgc ~ "Black globe temperature",
#                            SRWm2 ~ "Solar radiation in Watts",
#                            DP ~ "Dew Point"),
#               statistic = all_continuous() ~ "{mean} ({sd})",
#               missing = "ifany",
#               missing_text = "Missing") %>%
#   modify_spanning_header(update =  all_stat_cols() ~  "**Race**") %>%
#   modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
#   bold_labels()
#
# tbl_summary_female %>%
#   as_gt() %>%
#   cols_align(align = "center", columns = everything()) %>%
#   tab_options(
#     table.width = pct(100),
#     table.align = "center"
#   ) %>%
#   cols_width(
#     everything() ~ px(75)
#   ) %>%
#   tab_header(
#     title = md("Table 1: Summary Statistics by Race")
#   ) %>%
#   tab_style(
#     style = cell_text(size = px(8)),
#     locations = cells_title(groups = "title")
#   ) %>%
#   tab_style(
#     style = cell_text(size = px(7)),
#     locations = cells_body()
#   ) %>%
#   tab_style(
#     style = cell_text(size = px(7)),
#     locations = cells_column_labels()
#   )
```

```
# tbl_summary_male <- data %>%
#   filter(sex == "Male") %>%
#   select(-c(year, sex, CR, age_group, date)) %>%
#   mutate(
#     race = case_when(race == "Boston" ~ "Boston Marathon",
#                      race == "Chicago" ~ "Chicago Marathon",
#                      race == "NYC" ~ "New York City Marathon",
#                      race == "Twin Cities" ~ "Twin Cities Marathon",
#                      race == "Grandmas" ~ "Grandma's Marathon",
#                      TRUE ~ "Missing"),
#     flag = case_when(flag == 'White' ~ "WBGT < 10C",
#                      flag == 'Green' ~ "WBGT 10-18C",
#                      flag == 'Yellow' ~ "WBGT > 18-23C",
#                      flag == 'Red' ~ "WBGT > 23-28C",
#                      TRUE ~ "Missing")) %>%
#   tbl_summary(by = race,
#               label = list(age ~ "Age",
#                            CR_pct ~ "Percent off current course record ",
#                            Tdc ~ "Dry bulb temperature",
#                            Twc ~ "Wet bulb temperature",
#                            rh ~ "Percent relative humidity",
#                            Tgc ~ "Black globe temperature",
#                            SRWm2 ~ "Solar radiation in Watts",
#                            DP ~ "Dew Point"),
#               statistic = all_continuous() ~ "{mean} ({sd})",
#               missing = "ifany",
#               missing_text = "Missing") %>%
#   modify_spanning_header(update =  all_stat_cols() ~  "**Race**") %>%
#   modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
#   bold_labels()
#
# tbl_summary_male %>%
#   as_gt() %>%
#   cols_align(align = "center", columns = everything()) %>%
#   tab_options(
#     table.width = pct(100),
#     table.align = "center"
#   ) %>%
#   cols_width(
#     everything() ~ px(75)
#   ) %>%
#   tab_header(
#     title = md("Table 1: Summary Statistics by Race")
#   ) %>%
#   tab_style(
#     style = cell_text(size = px(8)),
#     locations = cells_title(groups = "title")
#   ) %>%
#   tab_style(
#     style = cell_text(size = px(7)),
#     locations = cells_body()
#   ) %>%
#   tab_style(
```

```r
#     style = cell_text(size = px(7)),
#     locations = cells_column_labels()
#   )

# Correlation Plot
cor_matrix <- cor(data[,-c(1, 2, 3, 4, 5, 15, 16, 18)], use = "complete.obs")

corrplot(cor_matrix, method = "color", type = "lower",
         tl.col = "black", tl.cex = 0.8, addCoef.col = "black",
         number.cex = 0.7, col = colorRampPalette(c("steelblue", "white", "steelblue"))(200))
fig_Tdc <- ggplot(data, aes(x = Tdc, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "Tdc", x = "Temperature (Tdc)", y = "Percent Off Course Record",
       color = "Age Group", linetype = "Gender") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title = element_text(size = 8),
        plot.title = element_text(size = 10, face = "plain"),
        legend.text = element_text(size = 6),
        legend.title = element_text(size = 7),
        legend.key.width = unit(0.5, "cm"),
        legend.spacing.x = unit(0.2, "cm")) +
  guides(color = guide_legend(nrow = 1),
         linetype = guide_legend(nrow = 1))

# Repeat the same for Twc, Tgc, and DP plots
fig_Twc <- ggplot(data, aes(x = Twc, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "Twc", x = "Temperature (Twc)", y = "Percent Off Course Record",
       color = "Age Group", linetype = "Gender") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title = element_text(size = 8),
        plot.title = element_text(size = 10, face = "plain"),
        legend.text = element_text(size = 6),
        legend.title = element_text(size = 7),
        legend.key.width = unit(0.5, "cm"),
        legend.spacing.x = unit(0.2, "cm")) +
  guides(color = guide_legend(nrow = 1),
         linetype = guide_legend(nrow = 1))

fig_Tgc <- ggplot(data, aes(x = Tgc, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "Tgc", x = "Temperature (Tgc)", y = "Percent Off Course Record",
       color = "Age Group", linetype = "Gender") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title = element_text(size = 8),
        plot.title = element_text(size = 10, face = "plain"),
        legend.text = element_text(size = 6),
        legend.title = element_text(size = 7),
        legend.key.width = unit(0.5, "cm"),
        legend.spacing.x = unit(0.2, "cm")) +
```

```
  guides(color = guide_legend(nrow = 1),
         linetype = guide_legend(nrow = 1))

fig_DP <- ggplot(data, aes(x = DP, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "DP", x = "Dew Point (DP)", y = "Percent Off Course Record",
       color = "Age Group", linetype = "Gender") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title = element_text(size = 8),
        plot.title = element_text(size = 10, face = "plain"),
        legend.text = element_text(size = 6),
        legend.title = element_text(size = 7),
        legend.key.width = unit(0.5, "cm"),
        legend.spacing.x = unit(0.2, "cm")) +
  guides(color = guide_legend(nrow = 1),
         linetype = guide_legend(nrow = 1))

# Arrange the plots with a more compact legend
temp_figs <- ggarrange(fig_Tdc, fig_Twc, fig_Tgc, fig_DP,
                       ncol = 2, nrow = 2,
                       common.legend = TRUE, legend = "bottom")

annotate_figure(temp_figs, top = text_grob("Overall Performance vs. Temperature Conditions",
                                           face = "bold", size = 12))

fig_WBGT <- ggplot(data, aes(x = WBGT, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  facet_wrap(~race) +
  labs(title = "WBGT",
       x = "WBGT", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 6),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

fig_SRWm2 <- ggplot(data, aes(x = SRWm2, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  facet_wrap(~race) +
  labs(title = "SRWm2",
       x = "Solar Radiation (SRWm2)", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 6),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))
```

```r
fig_rh <- ggplot(data, aes(x = rh, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  facet_wrap(~race) +
  labs(title = "rh",
       x = "Humidity (rh)", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 6),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

fig_wind <- ggplot(data, aes(x = Wind, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  facet_wrap(~race) +
  labs(title = "Wind",
       x = "Wind", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 6),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

fig_aqi <- ggplot(data, aes(x = aqi, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  facet_wrap(~race) +
  labs(title = "aqi",
       x = "Air Quality Index (aqi)", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 6),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

fig_WBGT
fig_SRWm2
fig_rh
fig_wind
fig_aqi
female_Tdc <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, Tdc) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE))

female_Tdc_fig <- ggplot(female_Tdc, aes(x = Tdc, y = mean_CR_pct, color = age_group)) +
```

```
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_Tdc <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, Tdc) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE))

male_Tdc_fig <- ggplot(male_Tdc, aes(x = Tdc, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_Tdc_fig, male_Tdc_fig, common.legend = TRUE)
female_Twc <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, Twc) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE))

female_Twc_fig <- ggplot(female_Twc, aes(x = Twc, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_Twc <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, Twc) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE))

male_Twc_fig <- ggplot(male_Twc, aes(x = Twc, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_Twc_fig, male_Twc_fig, common.legend = TRUE)
female_Tgc <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, Tgc) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_Tgc_fig <- ggplot(female_Tgc, aes(x = Tgc, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_Tgc <- data %>%
```

```
  filter(sex == "Male") %>%
  group_by(age_group, Tgc) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

male_Tgc_fig <- ggplot(male_Tgc, aes(x = Tgc, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_Tgc_fig, male_Tgc_fig, common.legend = TRUE)
low_humidity_df <- data %>%
  filter(rh < 5)

high_humidity_df <- data %>%
  filter(rh > 5)

female_low_rh <- low_humidity_df %>%
  filter(sex == "Female") %>%
  group_by(age_group, rh) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_low_rh_fig <- ggplot(female_low_rh, aes(x = rh, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female (Low Humidity)") +
  theme_minimal()

male_low_rh <- low_humidity_df %>%
  filter(sex == "Male") %>%
  group_by(age_group, rh) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

male_low_rh_fig <- ggplot(male_low_rh, aes(x = rh, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male (Low Humidity)") +
  theme_minimal()

female_high_rh <- high_humidity_df %>%
  filter(sex == "Female") %>%
  group_by(age_group, rh) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_high_rh_fig <- ggplot(female_high_rh, aes(x = rh, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female (High Humidity)") +
  theme_minimal()
```

```r
male_high_rh <- high_humidity_df %>%
  filter(sex == "Male") %>%
  group_by(age_group, rh) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

male_high_rh_fig <- ggplot(male_high_rh, aes(x = rh, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male (High Humidity)") +
  theme_minimal()

ggarrange(female_low_rh_fig, male_low_rh_fig, female_high_rh_fig, male_high_rh_fig, common.legend = TRU
female_SRWm2 <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, SRWm2) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_SRWm2_fig <- ggplot(female_SRWm2, aes(x = SRWm2, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_SRWm2 <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, SRWm2) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

male_SRWm2_fig <- ggplot(male_SRWm2, aes(x = SRWm2, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_SRWm2_fig, male_SRWm2_fig, common.legend = TRUE)
female_DP <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, DP) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_DP_fig <- ggplot(female_DP, aes(x = DP, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_DP <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, DP) %>%
```

```
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

male_DP_fig <- ggplot(male_DP, aes(x = DP, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_DP_fig, male_DP_fig, common.legend = TRUE)
female_aqi <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, aqi) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_aqi_fig <- ggplot(female_aqi, aes(x = aqi, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_aqi <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, aqi) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

male_aqi_fig <- ggplot(male_aqi, aes(x = aqi, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_aqi_fig, male_aqi_fig, common.legend = TRUE)
female_WBGT <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, WBGT) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()

female_WBGT_fig <- ggplot(female_WBGT, aes(x = WBGT, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_WBGT <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, WBGT) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE)) %>%
  ungroup()
```

```r
male_WBGT_fig <- ggplot(male_WBGT, aes(x = WBGT, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_WBGT_fig, male_WBGT_fig, common.legend = TRUE)
ggplot(data, aes(x = rh)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "black") +
  ggtitle("Distribution of Humidity Values") +
  theme_minimal() +
  labs(x = "Relative Humidity (%)", y = "Frequency")

low_humidity_df <- data %>%
  filter(rh < 5)

high_humidity_df <- data %>%
  filter(rh > 5)

by(data$age, data$sex, summary)
female_wind <- data %>%
  filter(sex == "Female") %>%
  group_by(age_group, Wind) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE))

female_wind_fig <- ggplot(female_wind, aes(x = Wind, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Female") +
  theme_minimal()

male_wind <- data %>%
  filter(sex == "Male") %>%
  group_by(age_group, Wind) %>%
  mutate(mean_CR_pct = mean(CR_pct, na.rm = TRUE))

male_wind_fig <- ggplot(male_wind, aes(x = Wind, y = mean_CR_pct, color = age_group)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  ggtitle("Male") +
  theme_minimal()

ggarrange(female_wind_fig, male_wind_fig, common.legend = TRUE)
# Fit the linear model with interactions
model_Tdc <- lm(CR_pct ~ Tdc*age + Tdc*sex, data = data)
summary(model_Tdc)

model_Twc <- lm(CR_pct ~ Tdc*age + Tdc*sex + Twc*age + Twc*sex, data = data)
summary(model_Twc)


model1 <- lmerTest::lmer(CR_pct ~ sex + rh + age + SRWm2 + Wind + WBGT + (1 | age), data = data)
summary(model1)
```