Insights into Marathon Performance: Impact of Weather on Age and Gender

William Qian

October 2024

Abstract

Introduction

Data Preprocessing

```
## [1] 11564    14
```

First, we will check for missing values and patterns in the data. We can easily find that there are some weather data missing in the dataset.

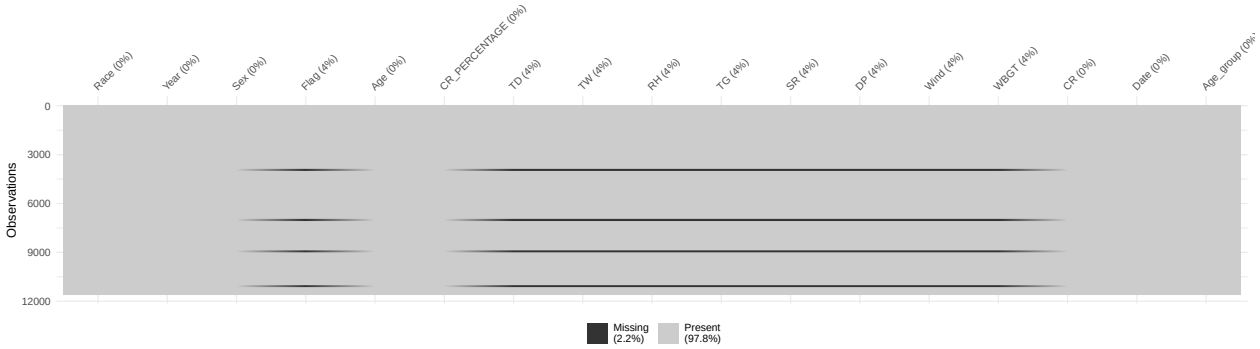
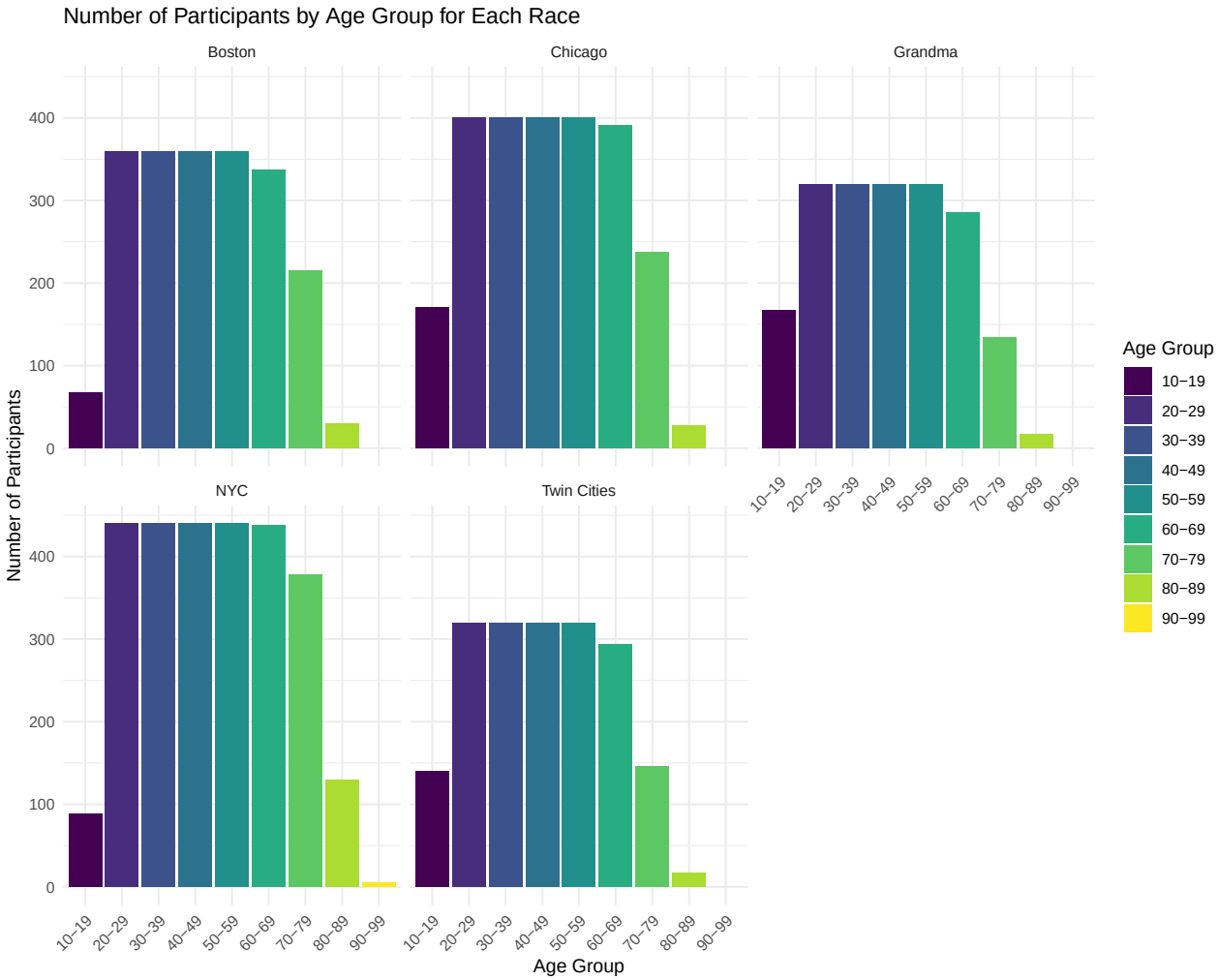


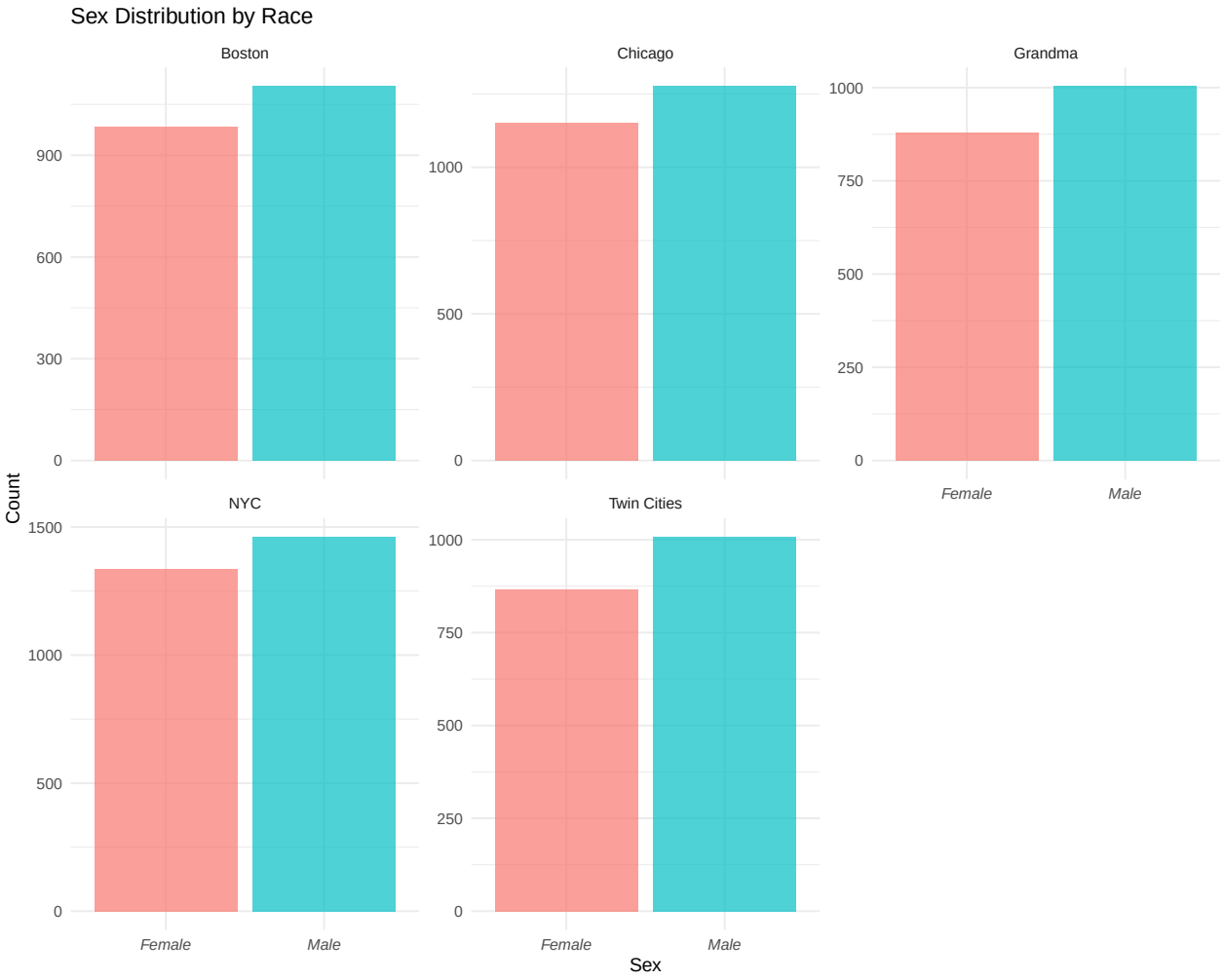
Table 1: Missing Percentage of Weather Data in Each Marathon by Year

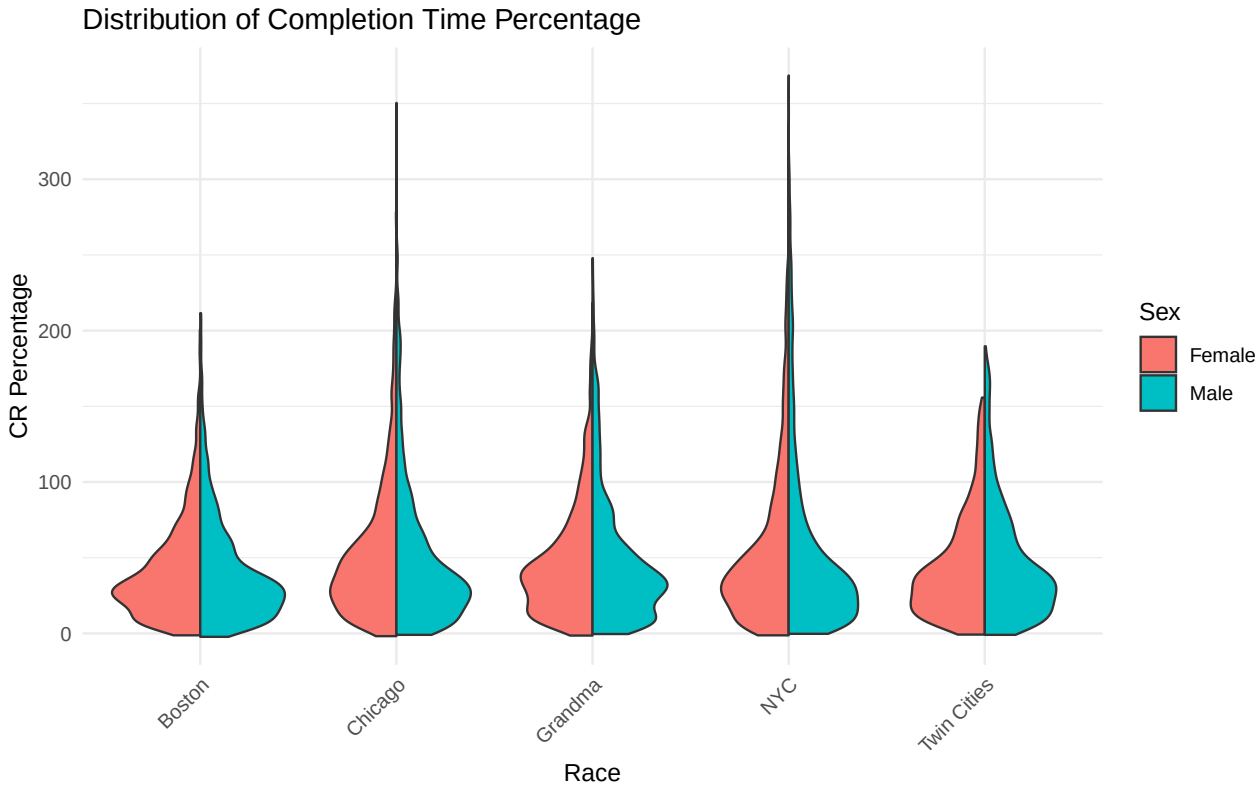
Year	Boston	Chicago	Grandma	NYC	Twin Cities
1993	0	0	NA	0	0
1994	0	0	NA	0	0
1995	0	0	NA	0	0
1996	0	0	NA	0	0
1997	0	0	NA	0	0
1998	0	0	NA	0	0
1999	0	0	NA	0	0
2000	0	0	0	0	0
2001	0	0	0	0	0
2002	0	0	0	0	0
2003	0	0	0	0	0
2004	0	0	0	0	0
2005	0	0	0	0	0

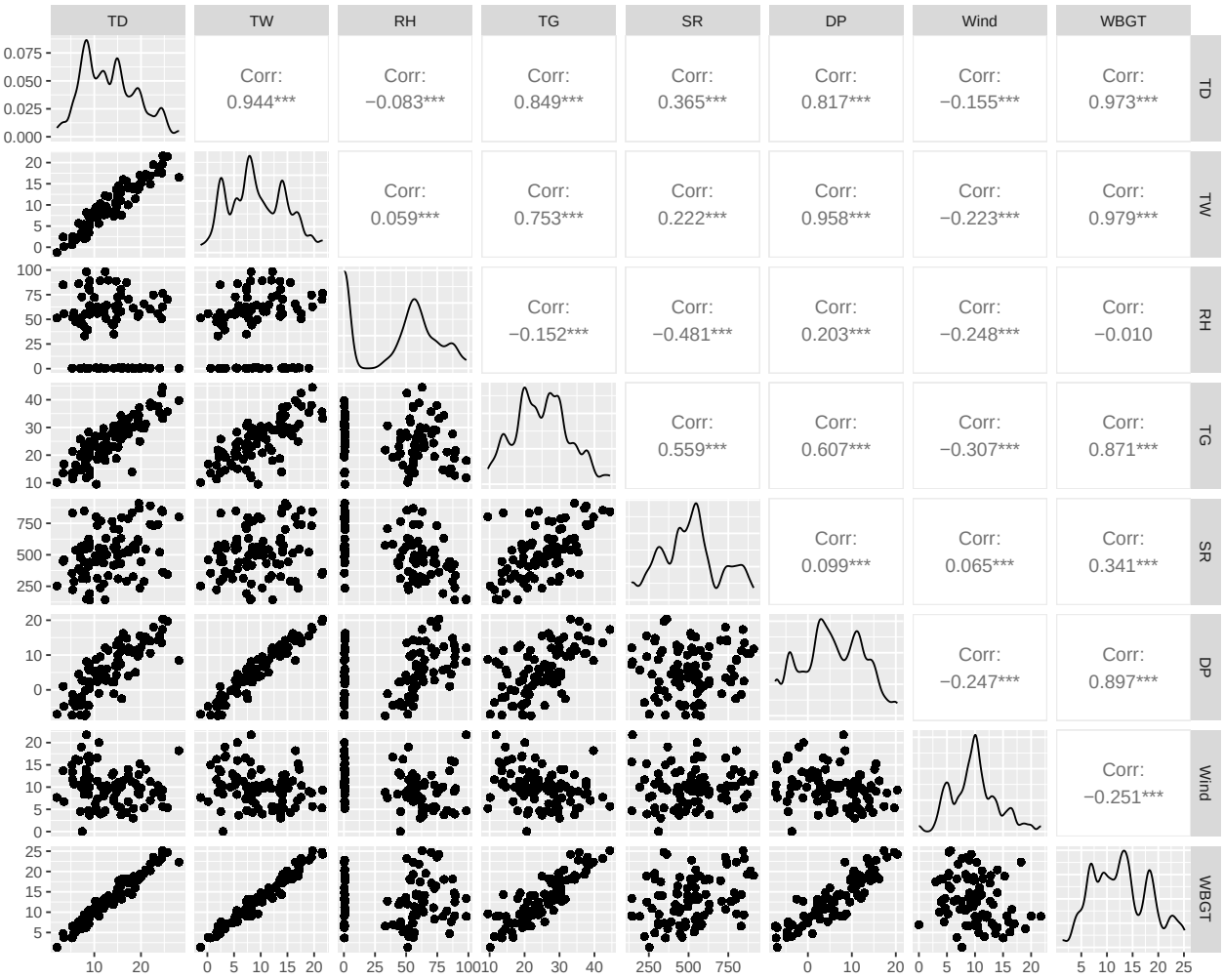
Year	Boston	Chicago	Grandma	NYC	Twin Cities
2006	0	0	0	0	0
2007	0	0	0	0	0
2008	0	0	0	0	0
2009	0	0	0	0	0
2010	0	0	0	0	0
2011	0	1	0	1	1
2012	0	0	1	0	0
2013	0	0	0	0	0
2014	0	0	0	0	0
2015	0	0	0	0	0
2016	0	0	0	0	0

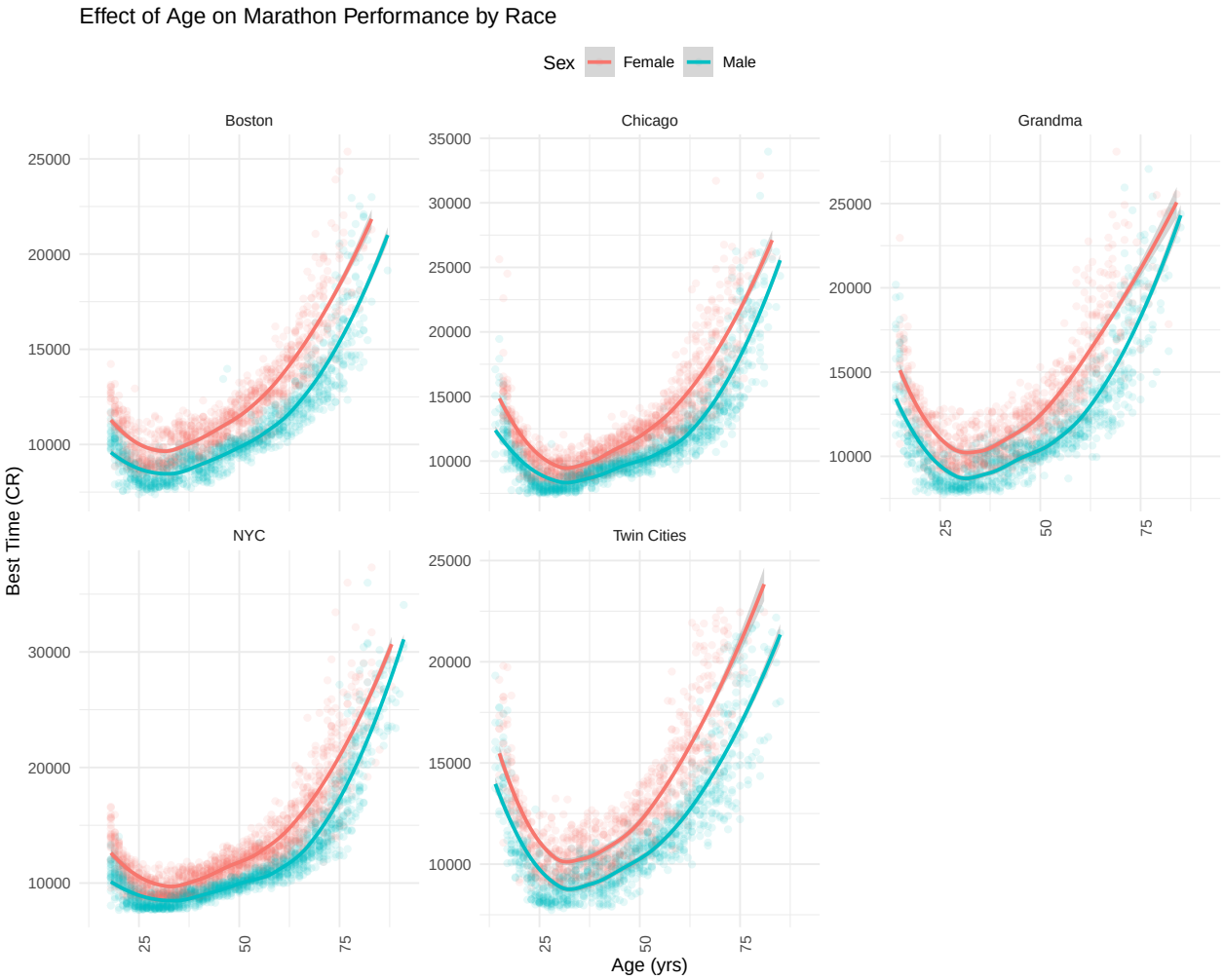
Data Analysis

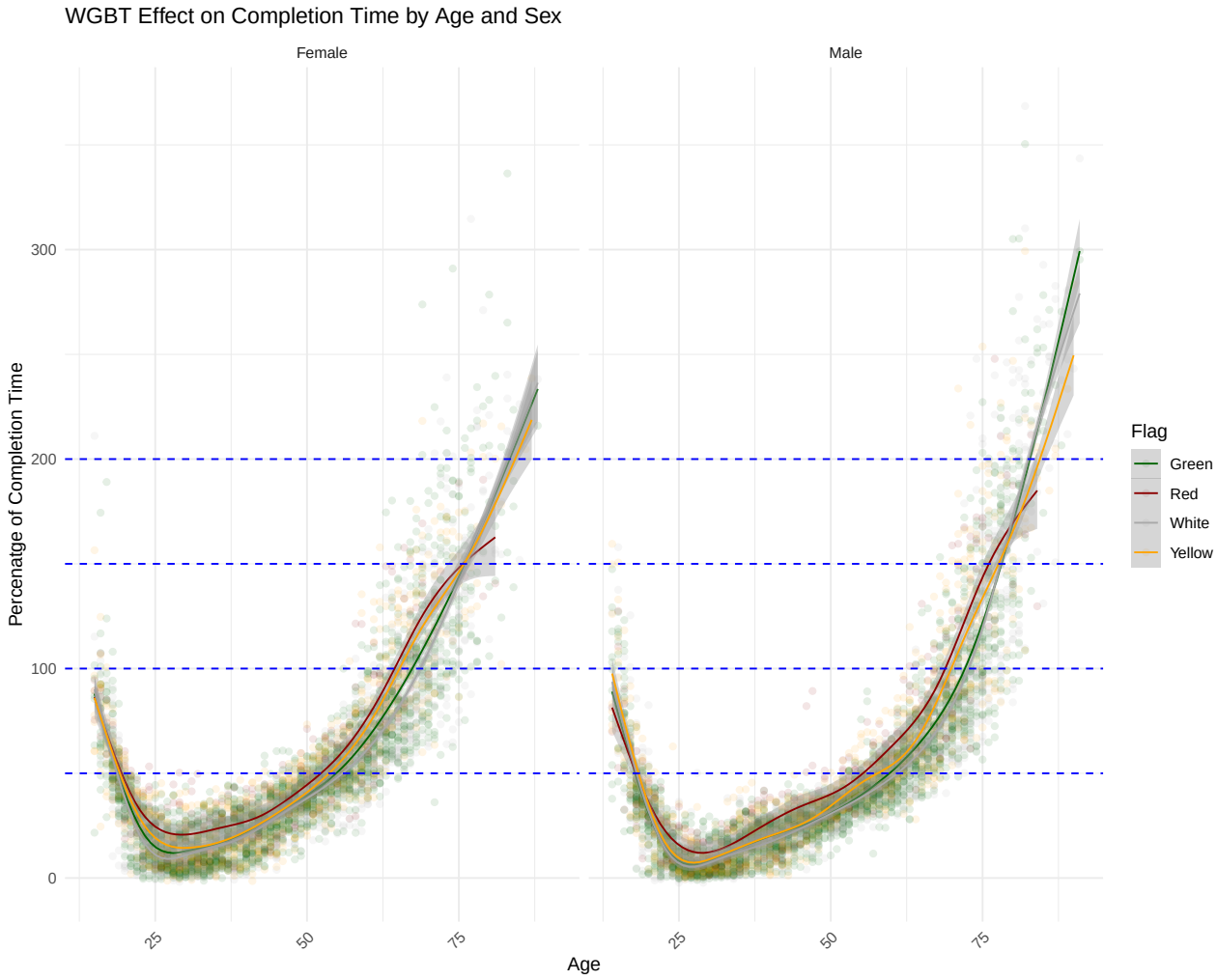



Sex Distribution by Race
Boston
Chicago
Grandma
NYC
Twin Cities
Count
Sex
Female
Male
0
300
600
900
500
1000
250
750
1500


Distribution of Completion Time Percentage
CR Percentage
300
200
100
0
Boston
Chicago
Grandma
NYC
Twin Cities
Race
Sex
Female
Male


TD
TW
RH
TG
SR
DP
Wind
WBGT
Corr: 0.944***
Corr: −0.083***
Corr: 0.849***
Corr: 0.365***
Corr: 0.817***
Corr: −0.155***
Corr: 0.973***
Corr: 0.059***
Corr: 0.753***
Corr: 0.222***
Corr: 0.958***
Corr: −0.223***
Corr: 0.979***
Corr: −0.152***
Corr: −0.481***
Corr: 0.203***
Corr: −0.248***
Corr: −0.010
Corr: 0.559***
Corr: 0.607***
Corr: −0.307***
Corr: 0.871***
Corr: 0.099***
Corr: 0.065***
Corr: 0.341***
Corr: −0.247***
Corr: 0.897***
Corr: −0.251***


Effect of Age on Marathon Performance by Race
Sex
Female
Male
Boston
Chicago
Grandma
NYC
Twin Cities
Best Time (CR)
Age (yrs)
25000
20000
15000
10000
35000
30000
25
50
75


WGBT Effect on Completion Time by Age and Sex
Female
Male
Percenatge of Completion Time
300
200
100
0
25
50
75
Age
Flag
Green
Red
White
Yellow

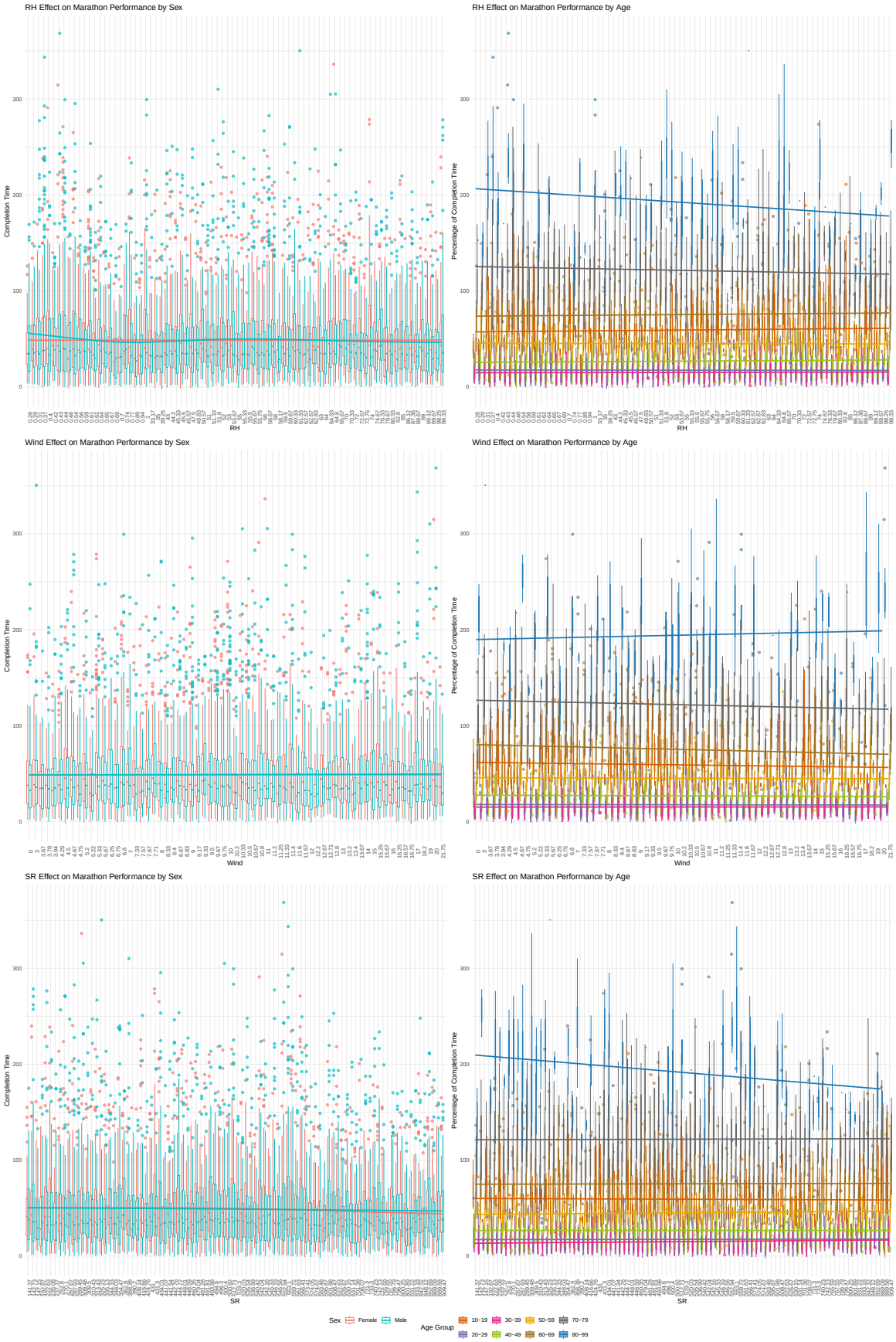

RH Effect on Marathon Performance by Sex
Completion Time
RH
RH Effect on Marathon Performance by Age
Percentage of Completion Time
RH
Wind Effect on Marathon Performance by Sex
Completion Time
Wind
Wind Effect on Marathon Performance by Age
Percentage of Completion Time
Wind
SR Effect on Marathon Performance by Sex
Completion Time
SR
SR Effect on Marathon Performance by Age
Percentage of Completion Time
SR
Sex
Female
Male
Age Group
10–19
20–29
30–39
40–49
50–59
60–69
70–79
80–99

Table 2: RH Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-30.2005520	0.9708892	-31.1060755	0.0000000
RH	-0.0003344	0.0094546	-0.0353719	0.9717838
SexMale	-4.7812957	0.6080433	-7.8634129	0.0000000
Age	1.7545477	0.0169002	103.8183237	0.0000000

Table 3: SR Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.4795524	1.2291307	-23.9840664	0.000000
SR	-0.0013813	0.0016146	-0.8555289	0.392277
SexMale	-4.7762996	0.6080477	-7.8551397	0.000000
Age	1.7539441	0.0169126	103.7061018	0.000000

Table 4: DP Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-32.017656	0.9181339	-34.872534	0
DP	0.287388	0.0435305	6.601993	0
SexMale	-4.801950	0.6068550	-7.912845	0
Age	1.759805	0.0168836	104.231898	0

Table 5: Wind Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28.5448183	1.1331262	-25.191208	0.0000000
Wind	-0.1734134	0.0743508	-2.332367	0.0196993
SexMale	-4.7828490	0.6078915	-7.867932	0.0000000
Age	1.7557419	0.0169015	103.880900	0.0000000

Table 6: Flag Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-30.666331	0.9422152	-32.547057	0.0000000
FlagRed	7.648126	1.3839258	5.526399	0.0000000
FlagWhite	-2.412088	0.6945525	-3.472866	0.0005169
FlagYellow	3.342042	0.8438010	3.960699	0.0000752
SexMale	-4.803419	0.6059713	-7.926810	0.0000000
Age	1.760180	0.0168528	104.444531	0.0000000

Table 7: WBGT Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-36.2759369	1.1404774	-31.807677	0
WBGT	0.4479189	0.0540060	8.293866	0
SexMale	-4.8069510	0.6061679	-7.930066	0
Age	1.7607921	0.0168625	104.420531	0

Table 8: Model with WBGT Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-35.8052408	2.1503858	-16.6506129	0.0000000
RH	-0.0100714	0.0117823	-0.8547937	0.3926839
SR	-0.0117774	0.0021134	-5.5728350	0.0000000
DP	-0.4666199	0.1192422	-3.9132102	0.0000916
Wind	0.0271575	0.0796785	0.3408382	0.7332318
WBGT	1.1062063	0.1550168	7.1360393	0.0000000
SexMale	-4.7708114	0.6053342	-7.8812856	0.0000000
Age	1.7557319	0.0168654	104.1026354	0.0000000

Table 9: Model with Flag Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-24.6502924	1.8792540	-13.1170628	0.0000000
RH	-0.0188363	0.0123629	-1.5236100	0.1276347
SR	-0.0089529	0.0019871	-4.5053890	0.0000067
DP	-0.1388574	0.0876138	-1.5848799	0.1130222
Wind	0.0582882	0.0842701	0.6916829	0.4891510
FlagRed	10.7582209	1.7404614	6.1812466	0.0000000
FlagWhite	-4.2883521	1.0310038	-4.1593950	0.0000322
FlagYellow	4.5640988	1.0933738	4.1743263	0.0000301
SexMale	-4.7755005	0.6055284	-7.8865016	0.0000000
Age	1.7556996	0.0168721	104.0591925	0.0000000

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
library(mice, warn.conflicts = FALSE)
library(naniar)
library(ggplot2)
library(dplyr)
library(readr)
library(tidyr)
library(readxl)
library(ggpubr)
library(gtsummary)
library(GGally)
```

```r
library(ggcorrplot)
library(knitr)
library(kableExtra)
library(lubridate)
library(patchwork)
library(introdataviz)

# Load data
marathon_data <- read.csv("../Data/project1.csv")
course_record <- read.csv("../Data/course_record.csv")
marathon_dates <- read.csv("../Data/marathon_dates.csv")

# rename the column names that are too long to follow.
colnames(marathon_data)[1] <- "Race"
colnames(marathon_data)[3] <- "Sex"
colnames(marathon_data)[5] <- "Age"
colnames(marathon_data)[6] <- "CR_PERCENTAGE"
colnames(marathon_data)[7] <- "TD"
colnames(marathon_data)[8] <- "TW"
colnames(marathon_data)[9] <- "RH"
colnames(marathon_data)[10] <- "TG"
colnames(marathon_data)[11] <- "SR"

# data type conversion
marathon_data$Year <- as.factor(marathon_data$Year)
marathon_data$Race <- as.factor(marathon_data$Race)
marathon_data$Sex <- as.factor(marathon_data$Sex)
marathon_data$Flag <- as.factor(marathon_data$Flag)

marathon_data$Flag[marathon_data$Flag == ""] <- NA

# Check the dimension of the data
dim(marathon_data)

# replace marathon name with code name in marathon_dates
marathon_dates$marathon[marathon_dates$marathon == "Boston"] <- 0
marathon_dates$marathon[marathon_dates$marathon == "Chicago"] <- 1
marathon_dates$marathon[marathon_dates$marathon == "NYC"] <- 2
marathon_dates$marathon[marathon_dates$marathon == "Twin Cities"] <- 3
marathon_dates$marathon[marathon_dates$marathon == "Grandmas"] <- 4
marathon_dates$marathon <- as.factor(marathon_dates$marathon)
colnames(marathon_dates)[1] <- "Race"

# rename date and year columns in marathon_dates
colnames(marathon_dates)[2] <- "Date"
colnames(marathon_dates)[3] <- "Year"

# replace marathon name with code name in course_record
course_record$Race[course_record$Race == "B"] <- 0
course_record$Race[course_record$Race == "C"] <- 1
course_record$Race[course_record$Race == "NY"] <- 2
course_record$Race[course_record$Race == "TC"] <- 3
course_record$Race[course_record$Race == "D"] <- 4
```

```r
course_record$Race <- as.factor(course_record$Race)

# replace gender in course_record
course_record$Gender[course_record$Gender == "M"] <- 1
course_record$Gender[course_record$Gender == "F"] <- 0
course_record$Gender <- as.factor(course_record$Gender)
colnames(course_record)[4] <- "Sex"

# Transform records in course_record into seconds
course_record$CR <- period_to_seconds(hms(course_record$CR))

# Join course_record and marathon_data
marathon_data <- merge(marathon_data, course_record, by = c("Race", "Year", "Sex"))

# Join marathon_data and marathon_dates
marathon_data <- merge(marathon_data, marathon_dates, by = c("Race", "Year"))

# calculate the record of each runner
marathon_data$CR <- (1 + marathon_data$CR_PERCENTAGE * 0.01) * marathon_data$CR

marathon_data <- marathon_data %>%
  mutate(Race = case_when(
    Race == 0 ~ "Boston",
    Race == 1 ~ "Chicago",
    Race == 2 ~ "NYC",
    Race == 3 ~ "Twin Cities",
    Race == 4 ~ "Grandma"
  ),
  Sex = case_when(
    Sex == 1 ~ "Male",
    Sex == 0 ~ "Female"
  )) %>%
  mutate(Age_group = cut(Age, breaks = seq(0, 100, by = 10), right = FALSE,
                         labels = c("0-9", "10-19", "20-29", "30-39", "40-49",
                                    "50-59", "60-69", "70-79", "80-89", "90-99")))

# Check for missing values and patterns
vis_miss(marathon_data)

# Check the missing percentage of weather data in each marathon by year
marathon_data %>%
  group_by(Race, Year) %>%
  summarise(missing_percentage = sum(is.na(Flag)) / n()) %>%
  pivot_wider(names_from="Race", values_from = missing_percentage) %>%
  arrange(Year) %>%
  replace_na(list(Boston = 0, Chicago = 0, NYC = 0, `Twin Cities` = 0, Grandmas = 0)) %>%
  kable(caption = "Missing Percentage of Weather Data in Each Marathon by Year")

# remove missing data
marathon_data <- marathon_data %>% filter(!is.na(Flag))

ggplot(marathon_data, aes(x = Age_group, fill = Age_group)) +
  geom_bar() +
```

```
  facet_wrap(~ Race) +
  scale_fill_viridis_d() +
  labs(title = "Number of Participants by Age Group for Each Race",
       x = "Age Group",
       y = "Number of Participants",
       fill = "Age Group") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
marathon_data <- marathon_data %>%
  mutate(Age_group = if_else(Age_group == "90-99", "80-99", Age_group)) %>%
  mutate(Age_group = if_else(Age_group == "80-89", "80-99", Age_group))
ggplot(marathon_data, aes(x = Sex, fill = Sex)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  facet_wrap(~ Race, scales = "free_y") +
  labs(title = "Sex Distribution by Race",
       x = "Sex",
       y = "Count",
       fill = "Sex") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(face = "italic"),
    legend.position = "none"
  )
ggplot(marathon_data, aes(x=Race, y = CR_PERCENTAGE, fill=Sex)) +
  geom_split_violin() +
  labs(title = "Distribution of Completion Time Percentage",
       y = "CR Percentage") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

ggpairs(marathon_data %>% select(TD, TW, RH, TG, SR, DP, Wind, WBGT))

ggplot(marathon_data, aes(x = Age, y = CR, color = Sex)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "loess", se = TRUE) +
  facet_wrap(~ Race, scales = "free_y") +
  labs(title = "Effect of Age on Marathon Performance by Race",
       x = "Age (yrs)",
       y = "Best Time (CR)",
       color = "Sex") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1),
    legend.position = "top"
  )
# WGBT effects
flag_colors <- c("Green" = "darkgreen",
                 "Yellow" = "orange",
                 "Red" = "darkred",
                 "White" = "darkgrey")
```

```r
marathon_data %>%
  ggplot(aes(x = Age, y = CR_PERCENTAGE, color = Flag)) +
  facet_wrap(~ Sex, scales = "fixed") +
  geom_point(alpha = 0.1) +
  geom_smooth(aes(group = Flag, color = Flag), se = T, size=0.5) +
  geom_hline(yintercept = 50, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 100, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 150, linetype = "dashed", color = "blue") +
  geom_hline(yintercept = 200, linetype = "dashed", color = "blue") +
  scale_color_manual(values = flag_colors) +
  labs(title = "WGBT Effect on Completion Time by Age and Sex",
       x = "Age",
       y = "Percenatge of Completion Time",
       color = "Flag") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

age_group_colors <- c("0-9" = "#1b9e77",
                      "10-19" = "#d95f02",
                      "20-29" = "#7570b3",
                      "30-39" = "#e7298a",
                      "40-49" = "#98c61e",
                      "50-59" = "#e6ab02",
                      "60-69" = "#a6761d",
                      "70-79" = "#666666",
                      "80-99" = "#1f78b4")

# RH Effects
RH_sex <- marathon_data %>%
      ggplot(aes(x = as.factor(round(RH, 2)), y = CR_PERCENTAGE, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = Sex), se = FALSE) +
      labs(title = "RH Effect on Marathon Performance by Sex",
           x = "RH",
           y = "Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

RH_age <- marathon_data %>%
      ggplot(aes(x = as.factor(round(RH, 2)), y = CR_PERCENTAGE, fill = Age_group, color = Age_group)) +
      geom_boxplot(alpha = 0.6, position = position_dodge(width = 1)) +
      geom_smooth(aes(group = Age_group, color = Age_group), method = "lm", se = FALSE) +
      scale_fill_manual(values = age_group_colors) +
      scale_color_manual(values = age_group_colors) +
      labs(title = "RH Effect on Marathon Performance by Age",
           x = "RH",
           y = "Percentage of Completion Time",
           fill = "Age Group",
           color = "Age Group") +
```

```
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

# Wind Effects
Wind_sex <- marathon_data %>%
      ggplot(aes(x = as.factor(round(Wind, 2)), y = CR_PERCENTAGE, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = Sex), se = FALSE) +
      labs(title = "Wind Effect on Marathon Performance by Sex",
           x = "Wind",
           y = "Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

Wind_age <- marathon_data %>%
      ggplot(aes(x = as.factor(round(Wind, 2)), y = CR_PERCENTAGE, fill = Age_group, color = Age_group))
      geom_boxplot(alpha = 0.6, position = position_dodge(width = 1)) +
      geom_smooth(aes(group = Age_group, color = Age_group), method = "lm", se = FALSE) +
      scale_fill_manual(values = age_group_colors) +
      scale_color_manual(values = age_group_colors) +
      labs(title = "Wind Effect on Marathon Performance by Age",
           x = "Wind",
           y = "Percentage of Completion Time",
           fill = "Age Group",
           color = "Age Group") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

# SR Effects
SR_sex <- marathon_data %>%
      ggplot(aes(x = as.factor(round(SR, 2)), y = CR_PERCENTAGE, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = Sex), se = FALSE) +
      labs(title = "SR Effect on Marathon Performance by Sex",
           x = "SR",
           y = "Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

SR_age <- marathon_data %>%
      ggplot(aes(x = as.factor(round(SR, 2)), y = CR_PERCENTAGE, fill = Age_group, color = Age_group)) +
      geom_boxplot(alpha = 0.6, position = position_dodge(width = 1)) +
      geom_smooth(aes(group = Age_group, color = Age_group), method = "lm", se = FALSE) +
      scale_fill_manual(values = age_group_colors) +
      scale_color_manual(values = age_group_colors) +
```

```
      labs(title = "SR Effect on Marathon Performance by Age",
           x = "SR",
           y = "Percentage of Completion Time",
           fill = "Age Group",
           color = "Age Group") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

(RH_sex | RH_age) / (Wind_sex | Wind_age) / (SR_sex | SR_age) +
  plot_layout(guides = "collect", axis_titles = "collect") &
  theme(legend.position = 'bottom')

# RH model
RH_model <- glm(CR_PERCENTAGE ~ RH + Sex + Age, data = marathon_data)
kable(summary(RH_model)$coefficients, caption = "RH Model Coefficients")

# SR model
SR_model <- glm(CR_PERCENTAGE ~ SR + Sex + Age, data = marathon_data)
kable(summary(SR_model)$coefficients, caption = "SR Model Coefficients")

# DP model
DP_model <- glm(CR_PERCENTAGE ~ DP + Sex + Age, data = marathon_data)
kable(summary(DP_model)$coefficients, caption = "DP Model Coefficients")

# Wind model
Wind_model <- glm(CR_PERCENTAGE ~ Wind + Sex + Age, data = marathon_data)
kable(summary(Wind_model)$coefficients, caption = "Wind Model Coefficients")

# Flag model
Flag_model <- glm(CR_PERCENTAGE ~ Flag + Sex + Age, data = marathon_data)
kable(summary(Flag_model)$coefficients, caption = "Flag Model Coefficients")

# WBGT model
WBGT_model <- glm(CR_PERCENTAGE ~ WBGT + Sex + Age, data = marathon_data)
kable(summary(WBGT_model)$coefficients, caption = "WBGT Model Coefficients")

# linear model
lm_model_1 <- glm(CR_PERCENTAGE ~ RH + SR + DP + Wind + WBGT + Sex + Age, data = marathon_data)
kable(summary(lm_model_1)$coefficients, caption = "Model with WBGT Coefficients")

lm_mode_2 <- glm(CR_PERCENTAGE ~ RH + SR + DP + Wind + Flag + Sex + Age, data = marathon_data)
kable(summary(lm_mode_2)$coefficients, caption = "Model with Flag Coefficients")
```