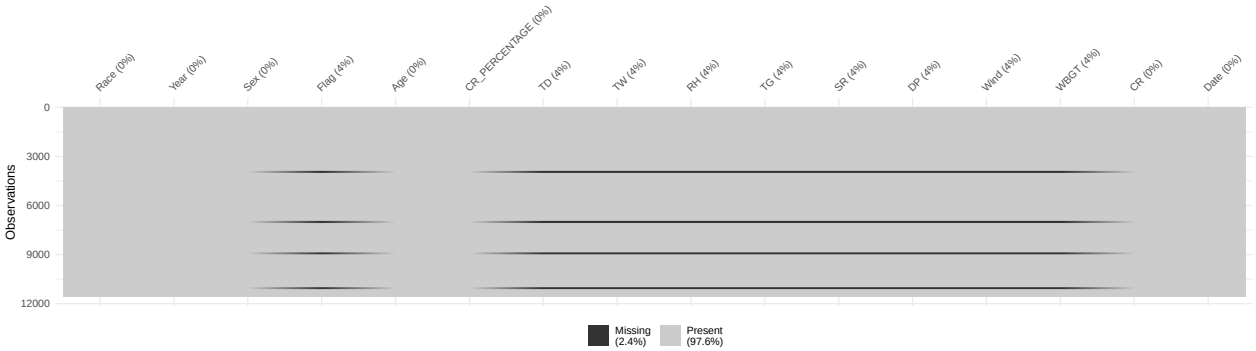
Abstract

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.

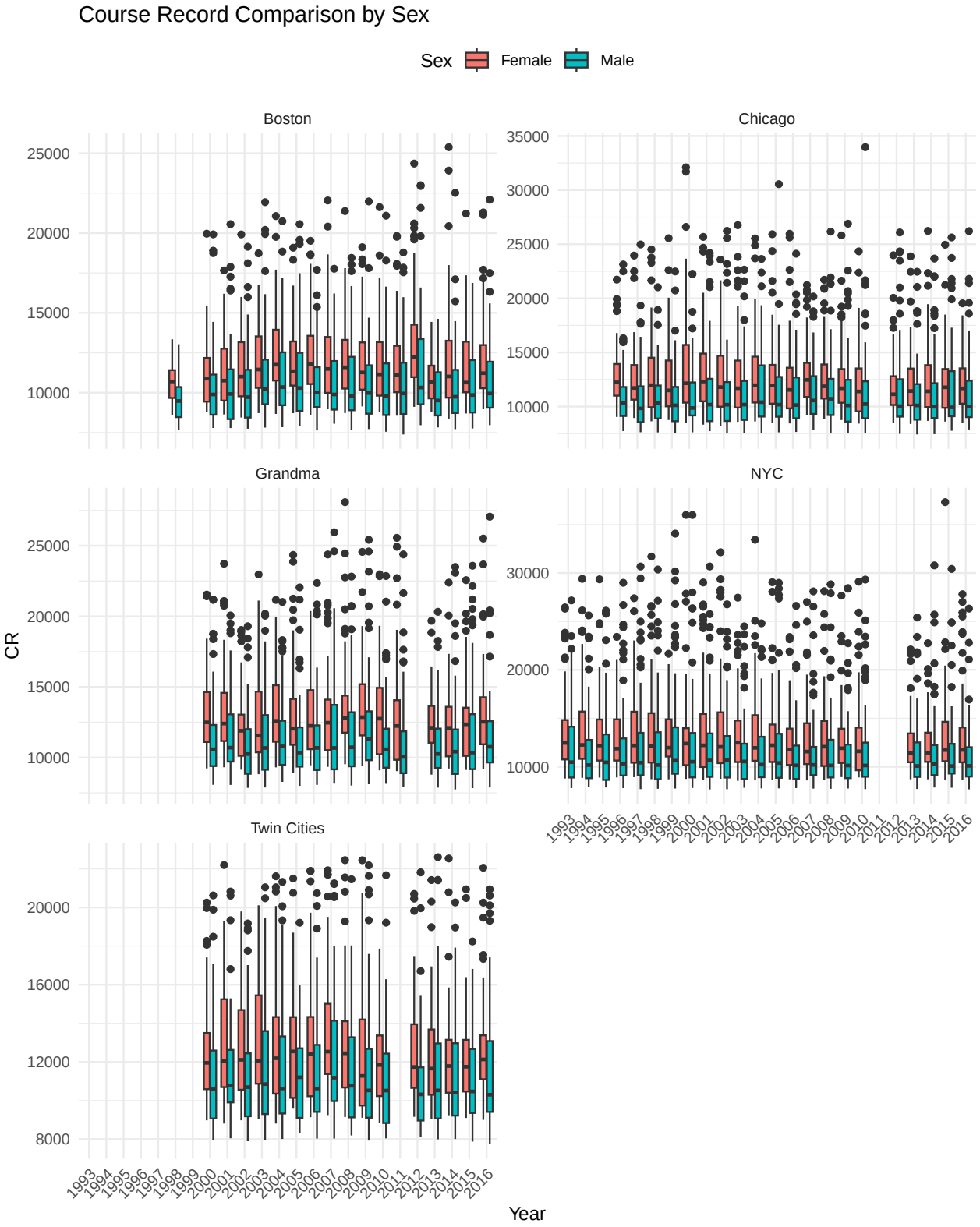



```
## `summarise()` has grouped output by 'Race'. You can override using the
## `.groups` argument.
```

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
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



Code Appendix

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

# Load data
marathon_data <- read.csv("../Data/project1.csv")
aqi_values <- read.csv("../Data/aqi_values.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)
```

```r
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
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"
  ))

# 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 = as.factor(Year), y = CR, fill = Sex)) +
  geom_boxplot() +
  facet_wrap(~ Race, scales = "free_y", ncol=2) +
  labs(title = "Course Record Comparison by Sex",
       x = "Year",
       y = "CR",
       fill = "Sex") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "top"
  )
```