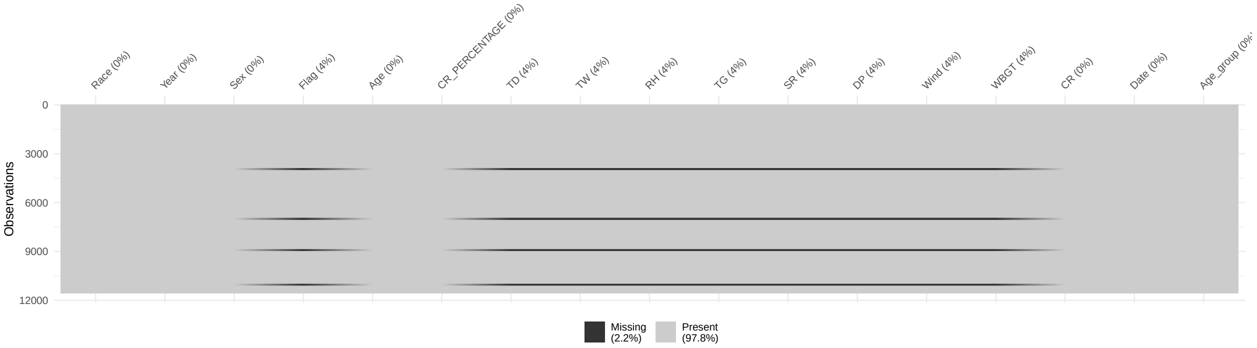
Abstract

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.

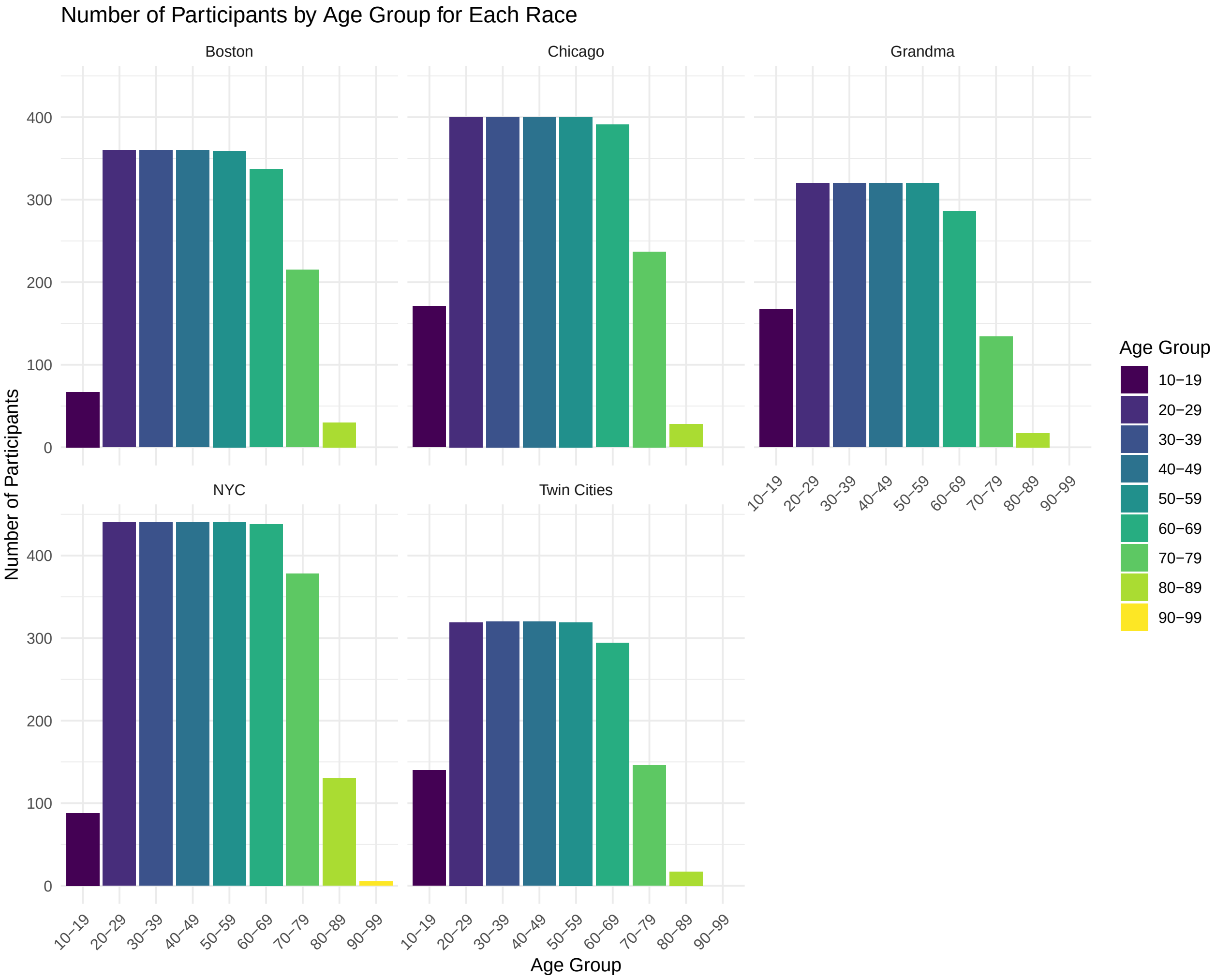


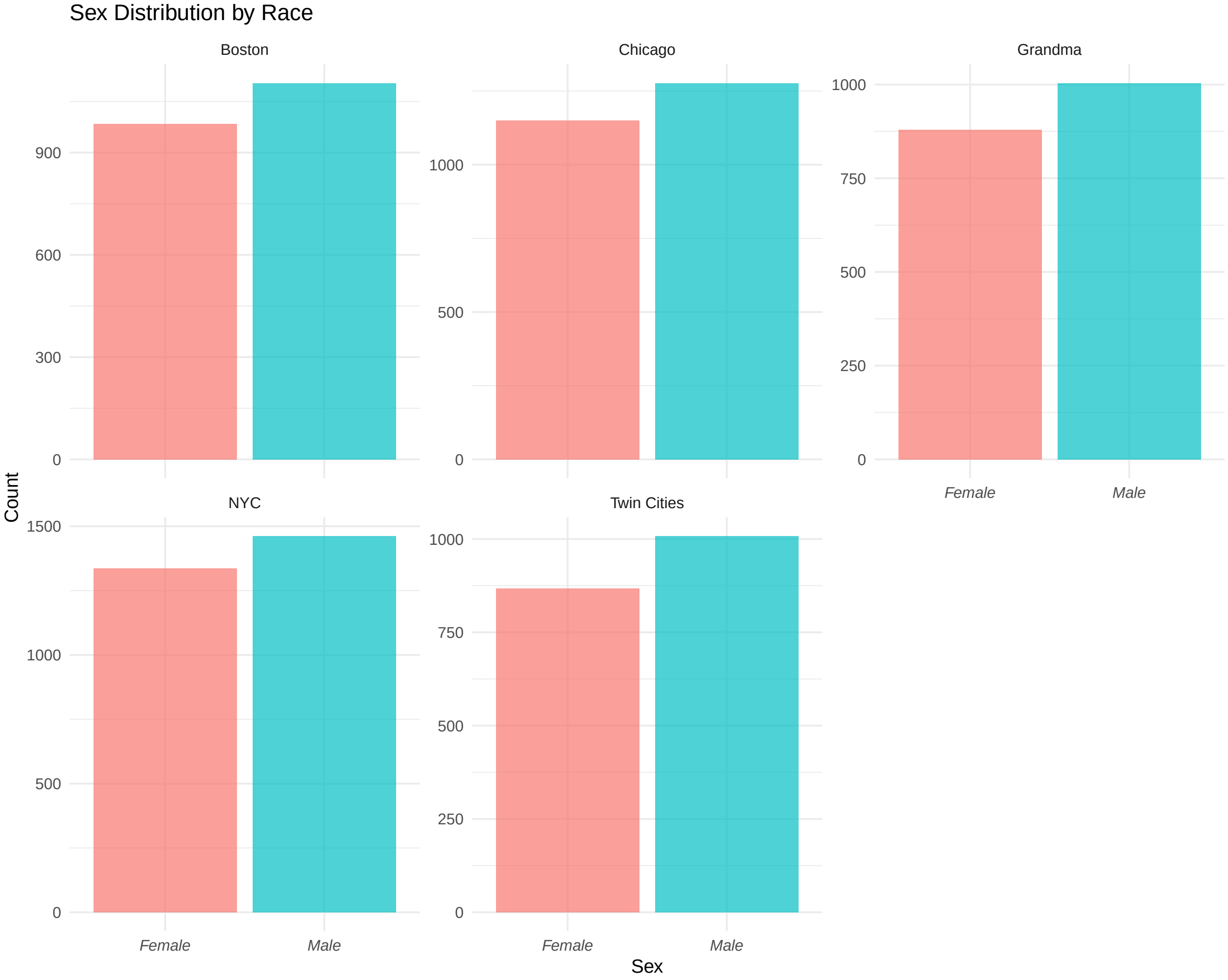
```
## `summarise()` has grouped output by 'Race'. You can override using the
## `.groups` argument.
```

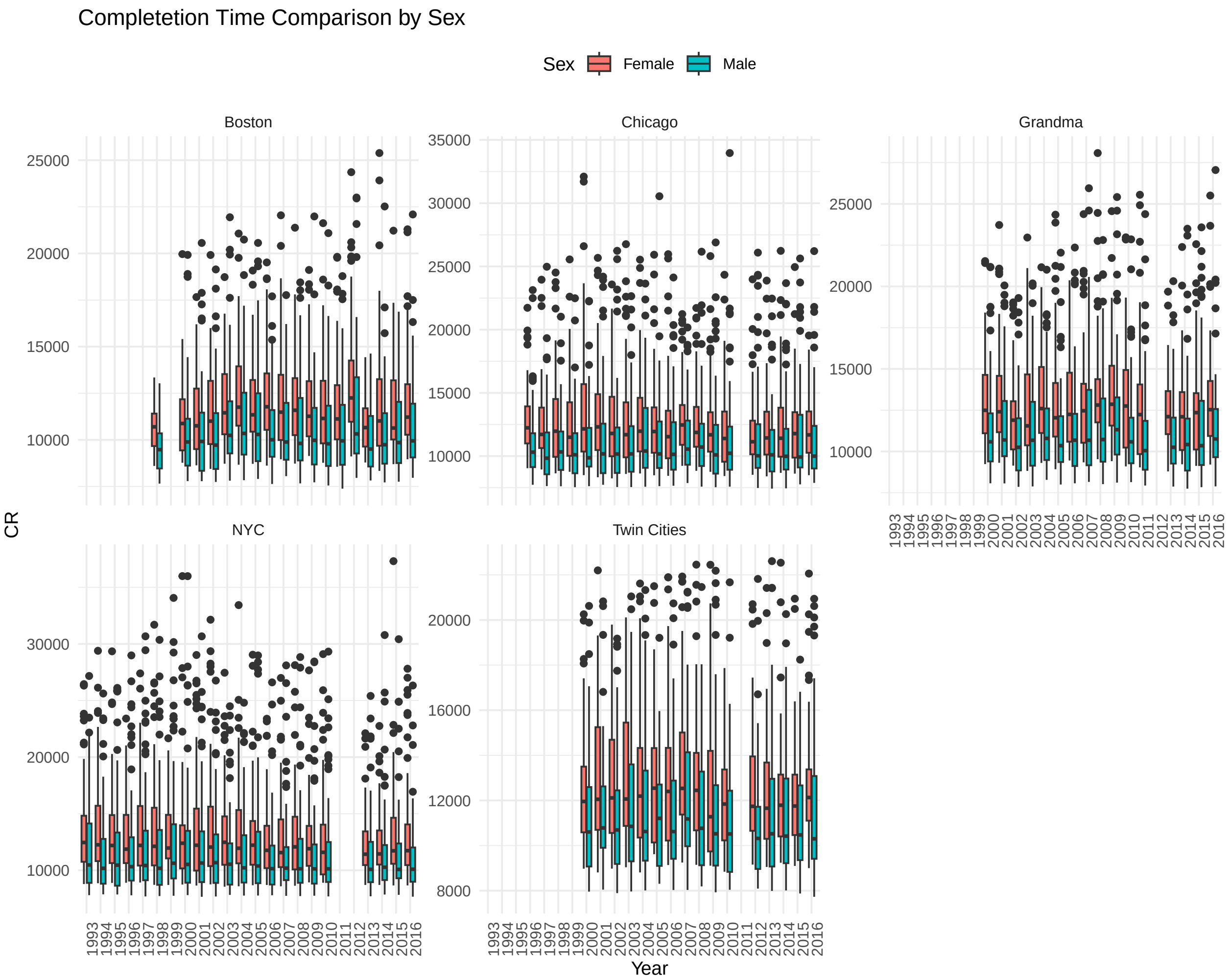
Table 1: Missing Percentage of Weather Data in Each Marathon by Year

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

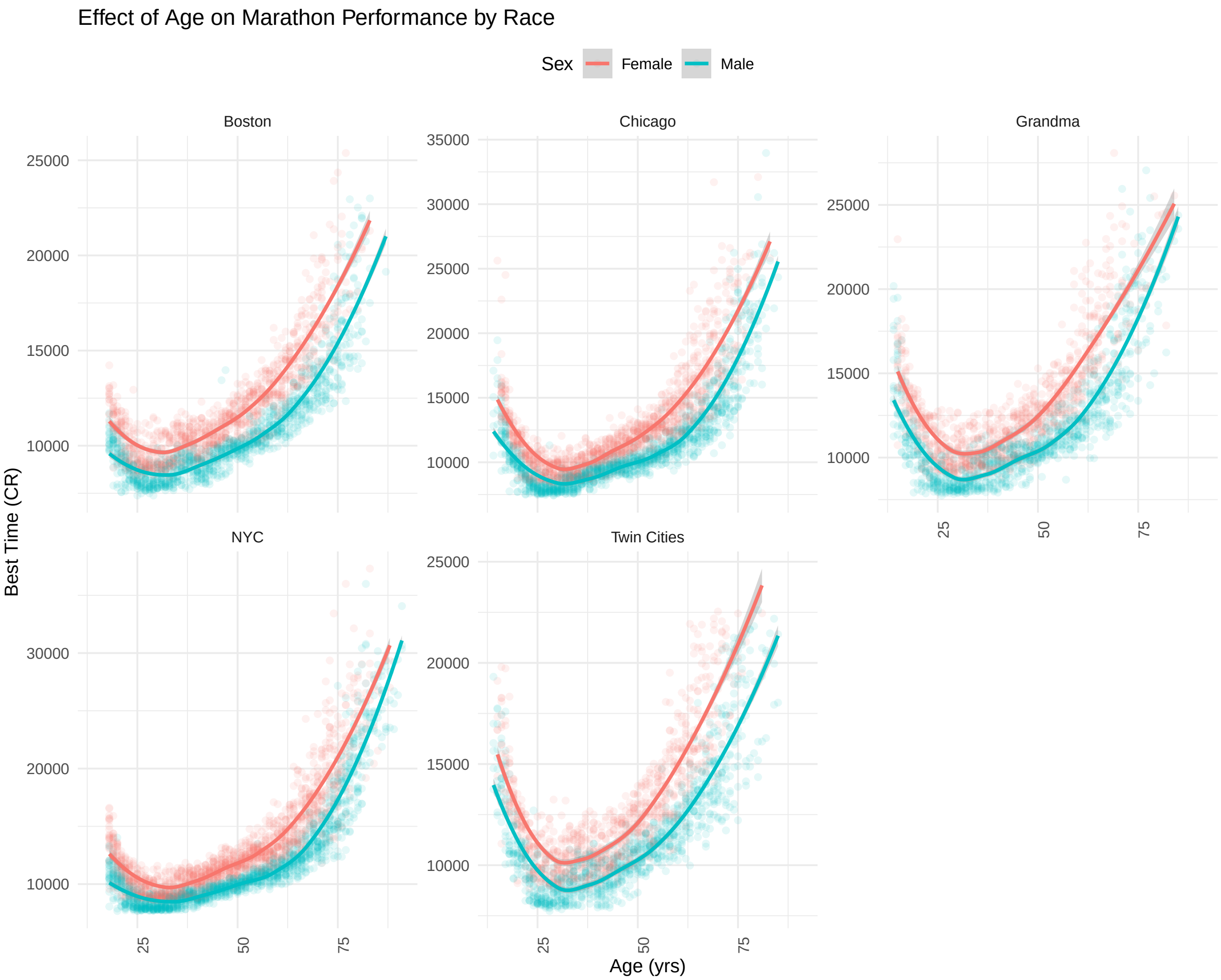
Data Analysis



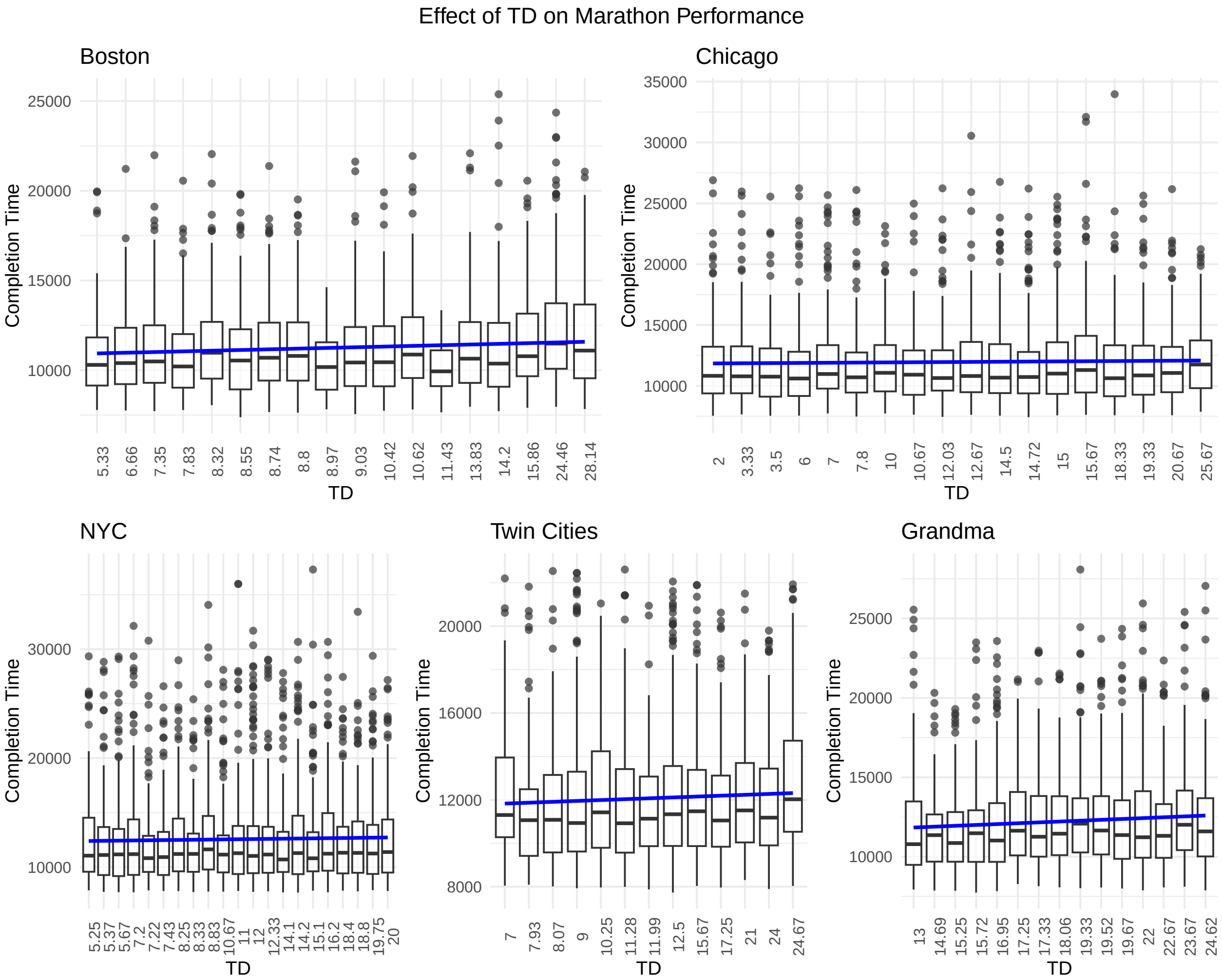

Sex Distribution by Race
Boston
Chicago
Grandma
NYC
Twin Cities
Count
Sex
Female
Male
0
300
600
900
0
500
1000
0
250
500
750
1000
0
500
1000
1500
0
250
500
750
1000



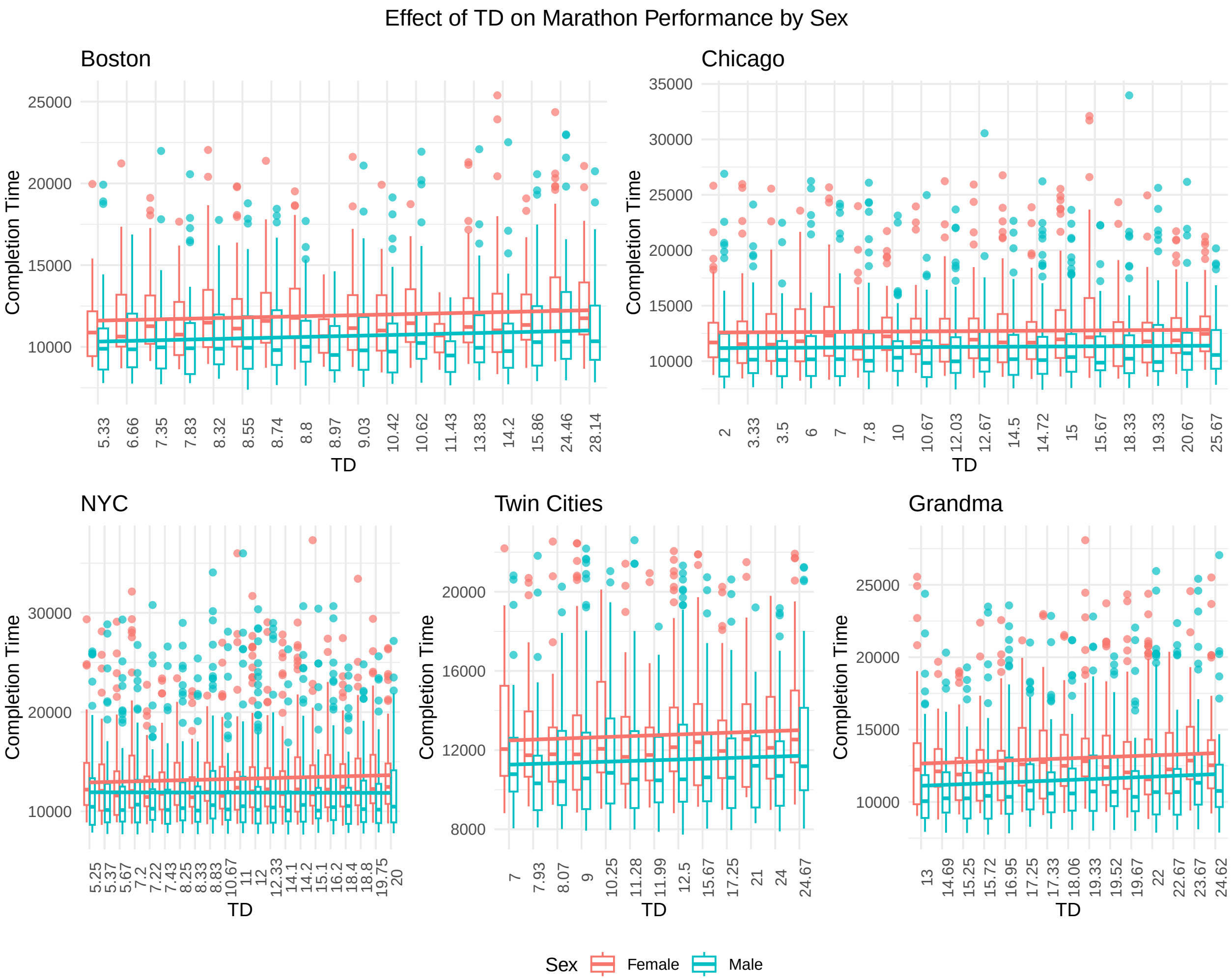
```
## `geom_smooth()` using formula = 'y ~ x'
```



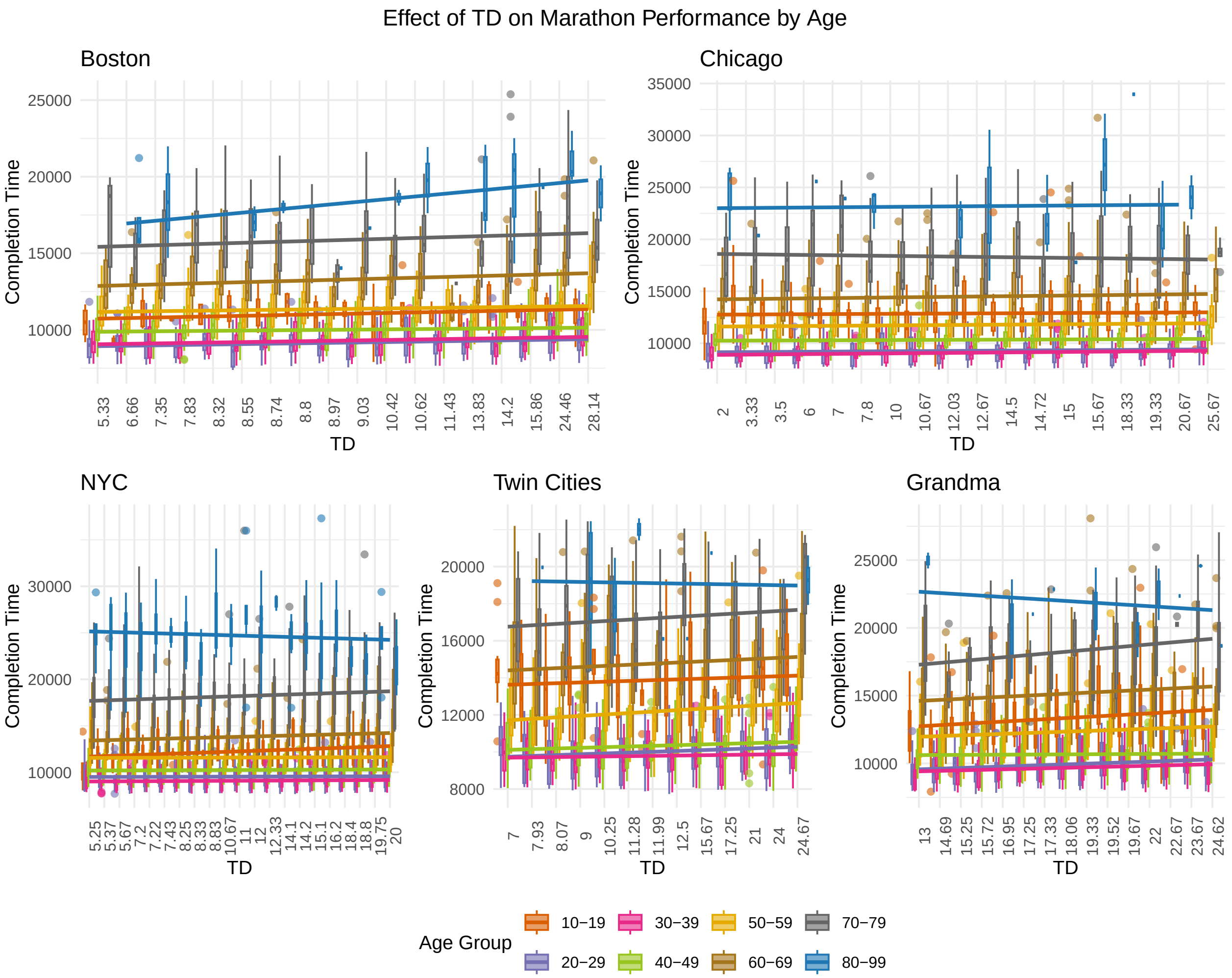
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



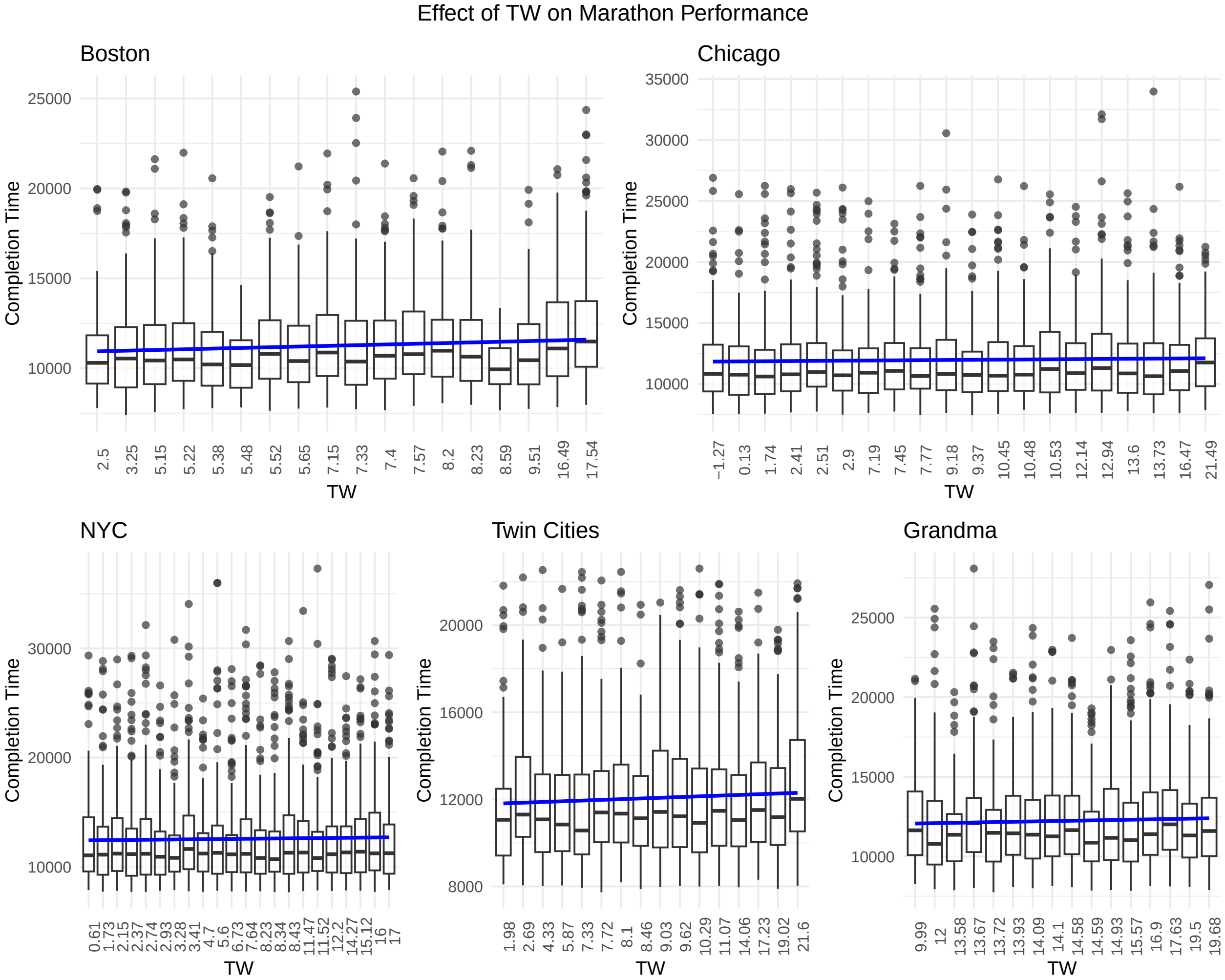
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



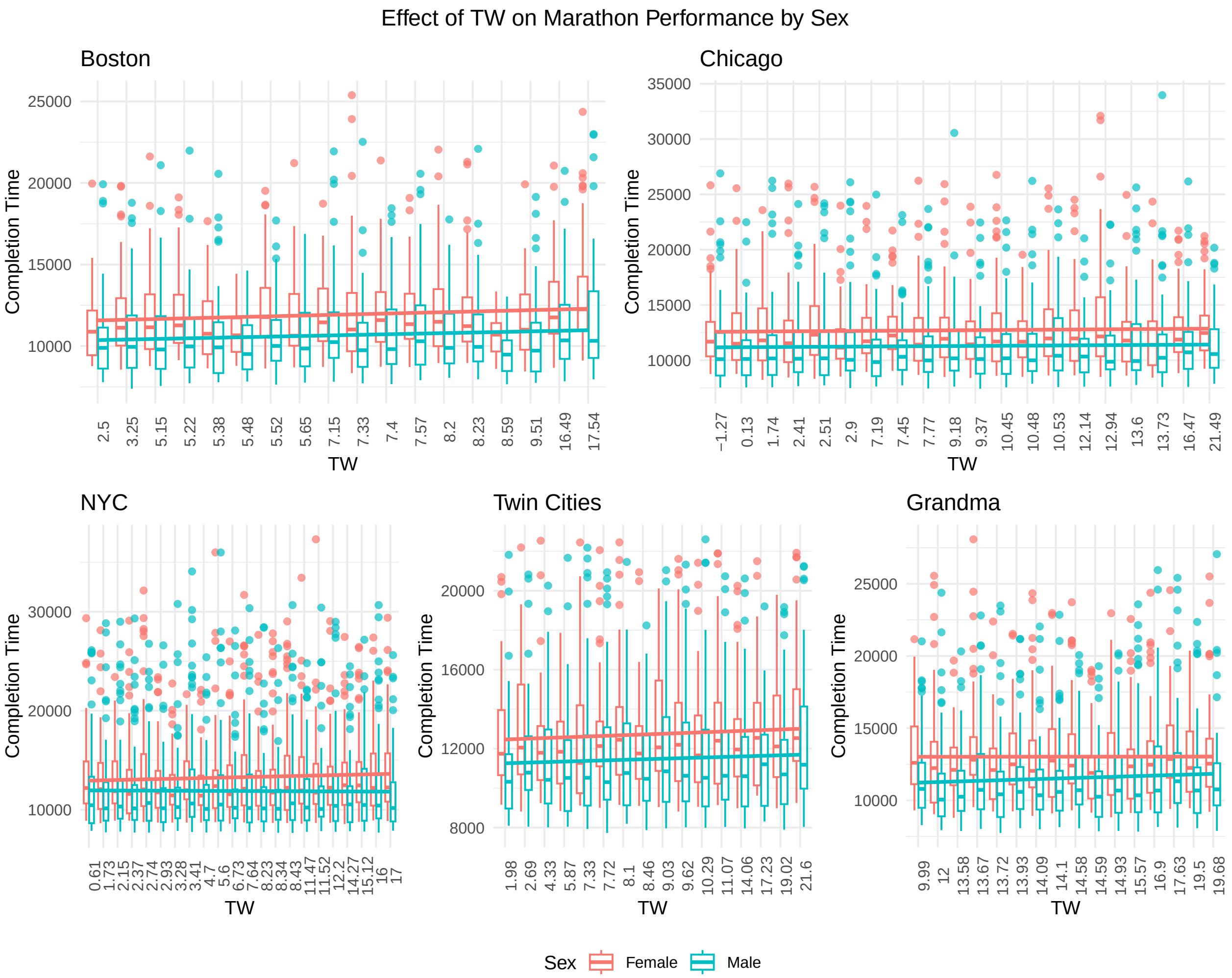
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



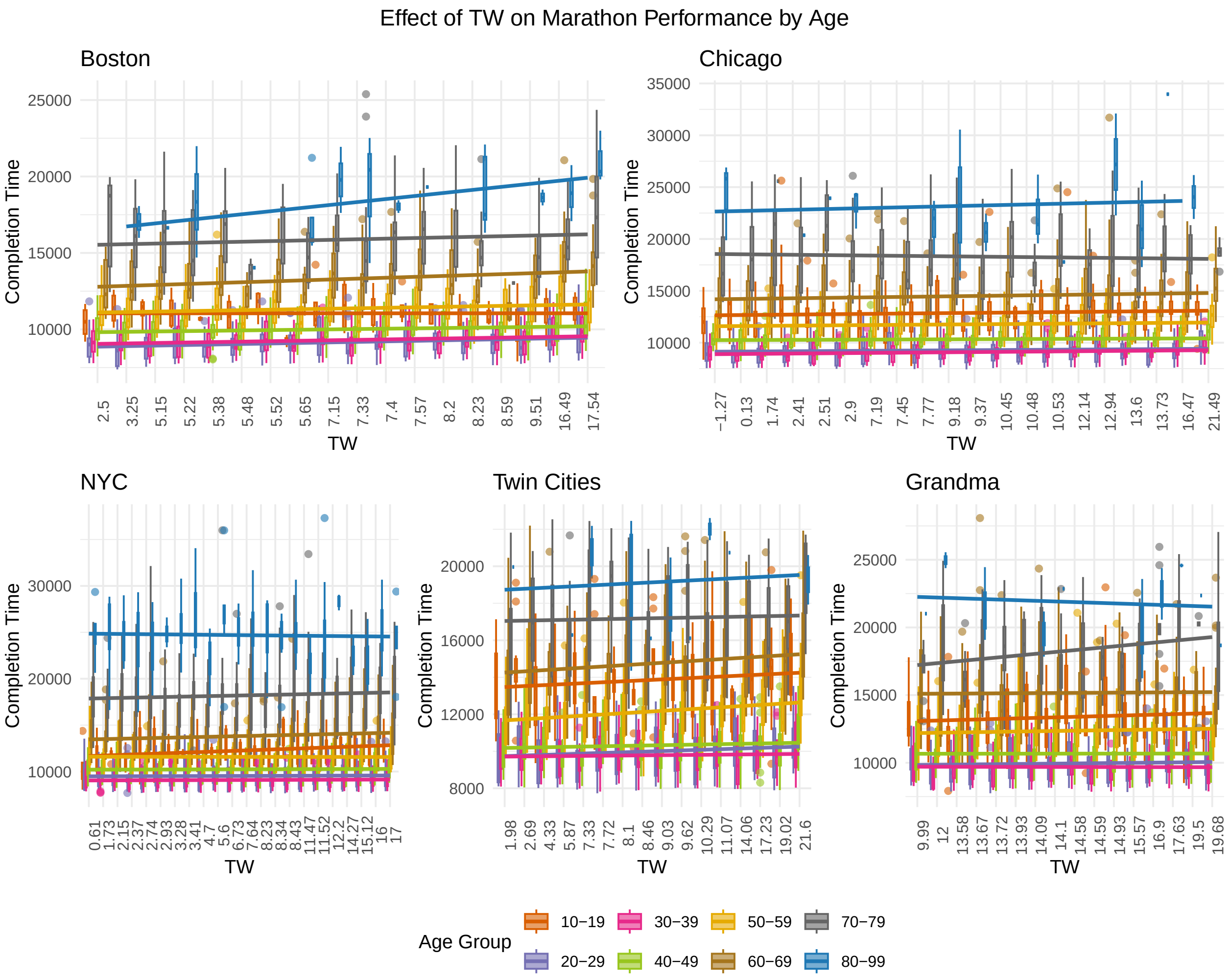
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



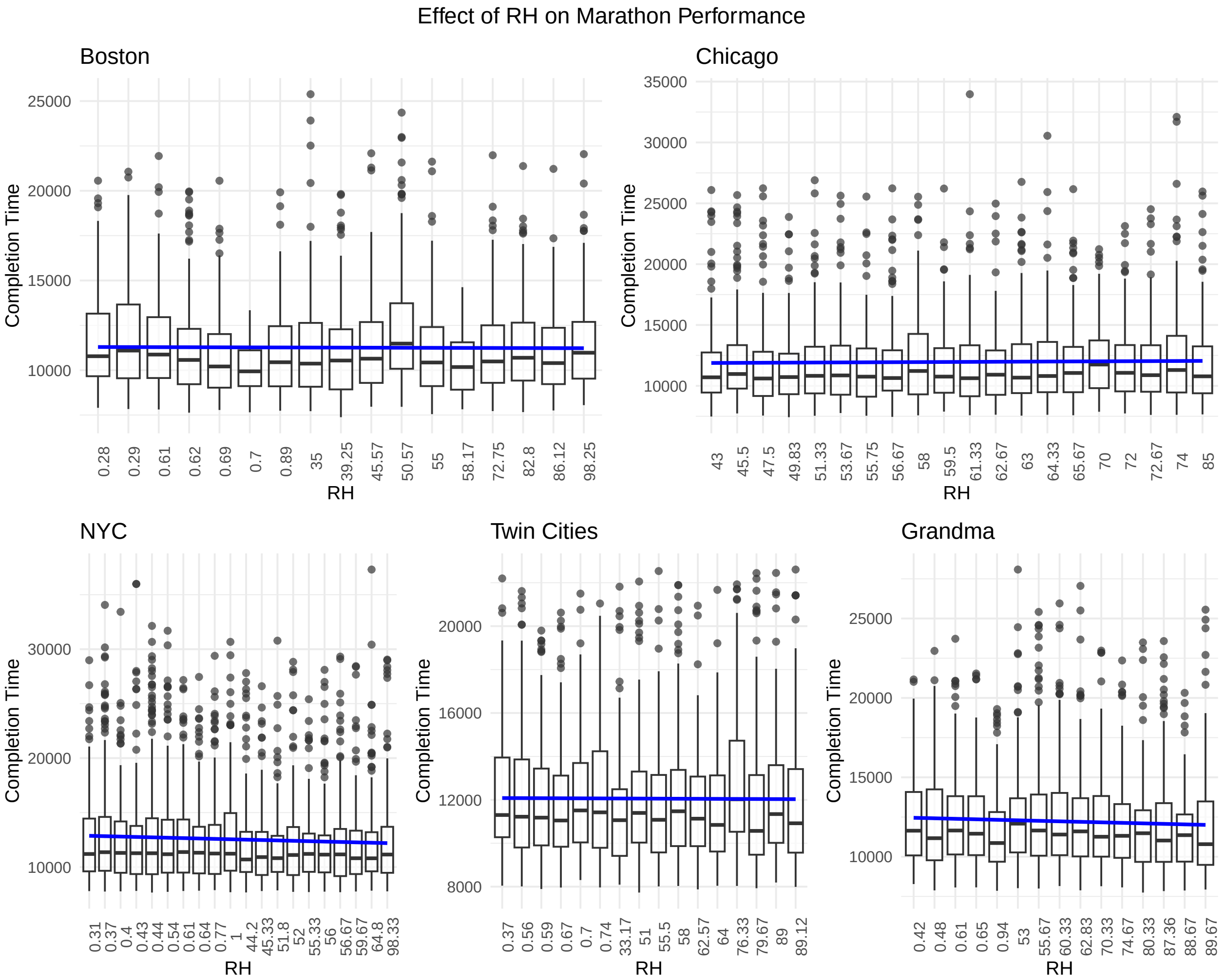
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



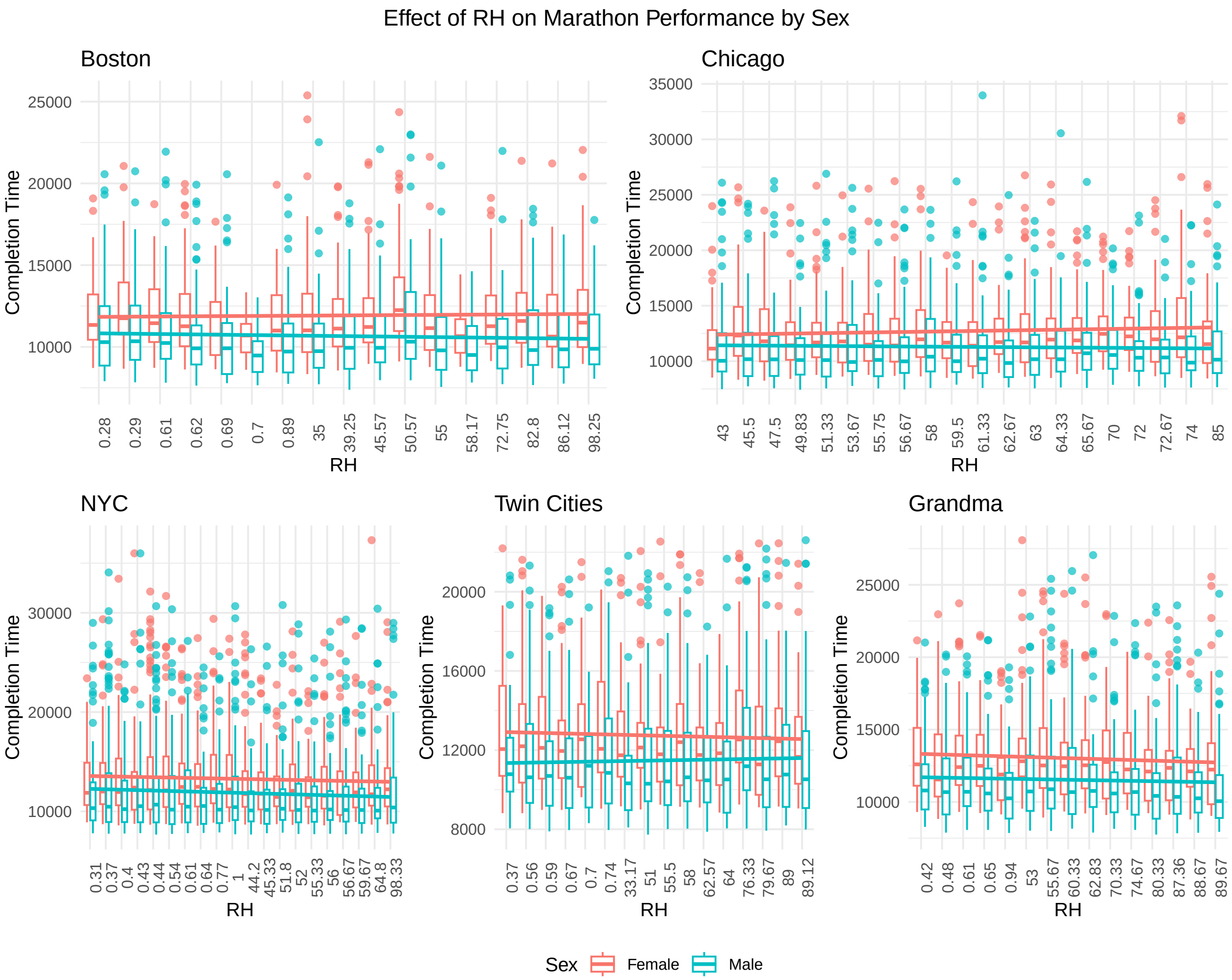
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



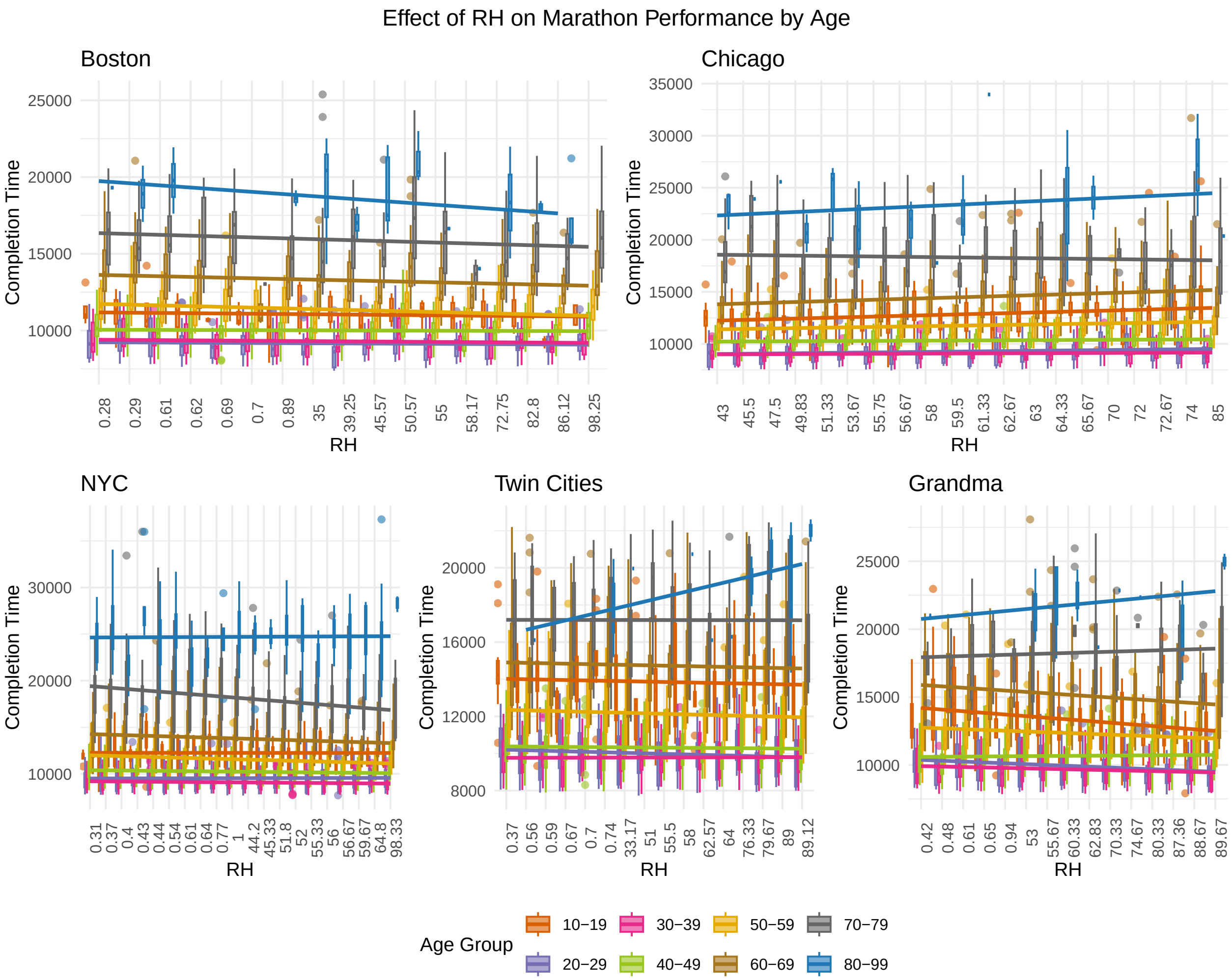
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



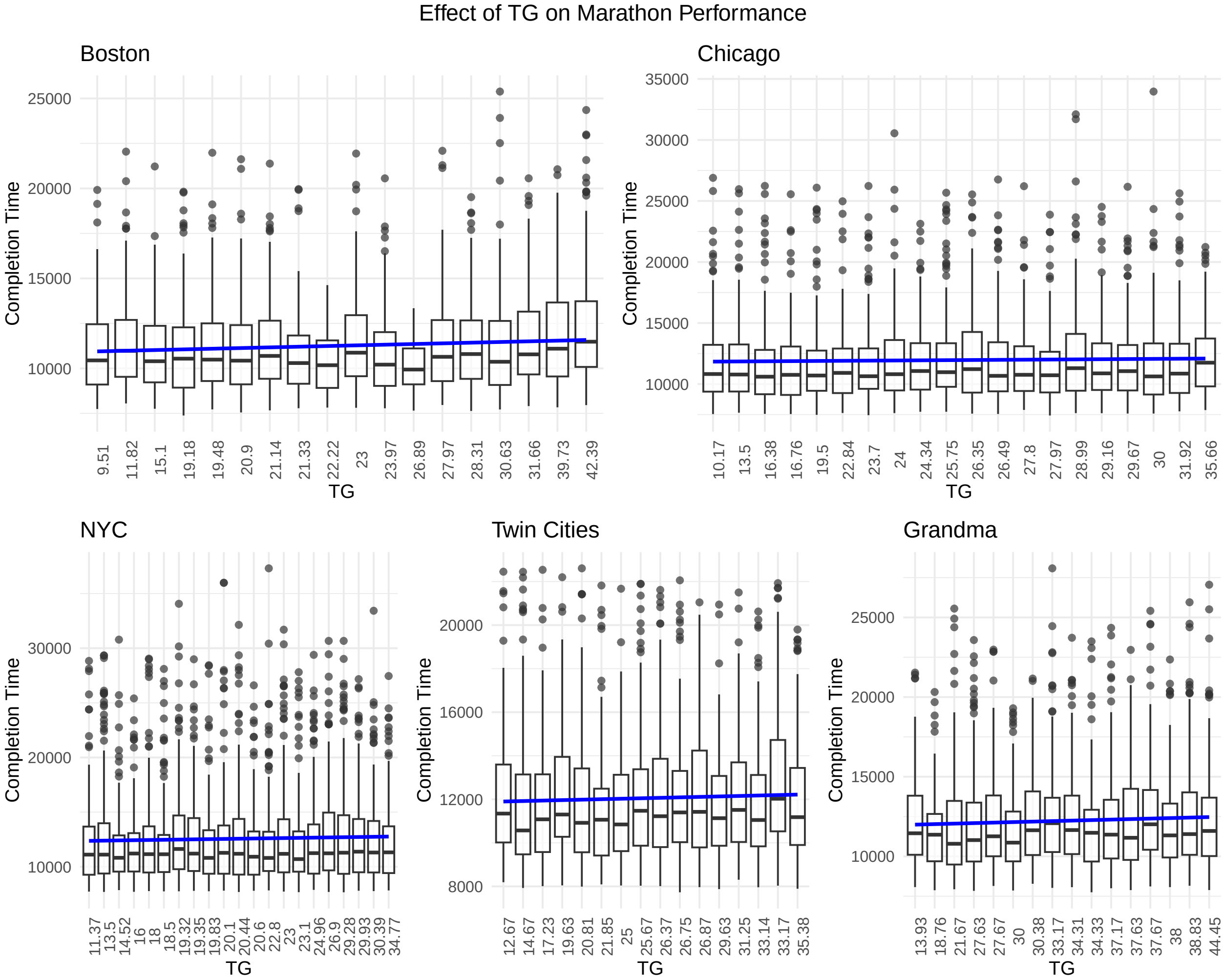
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



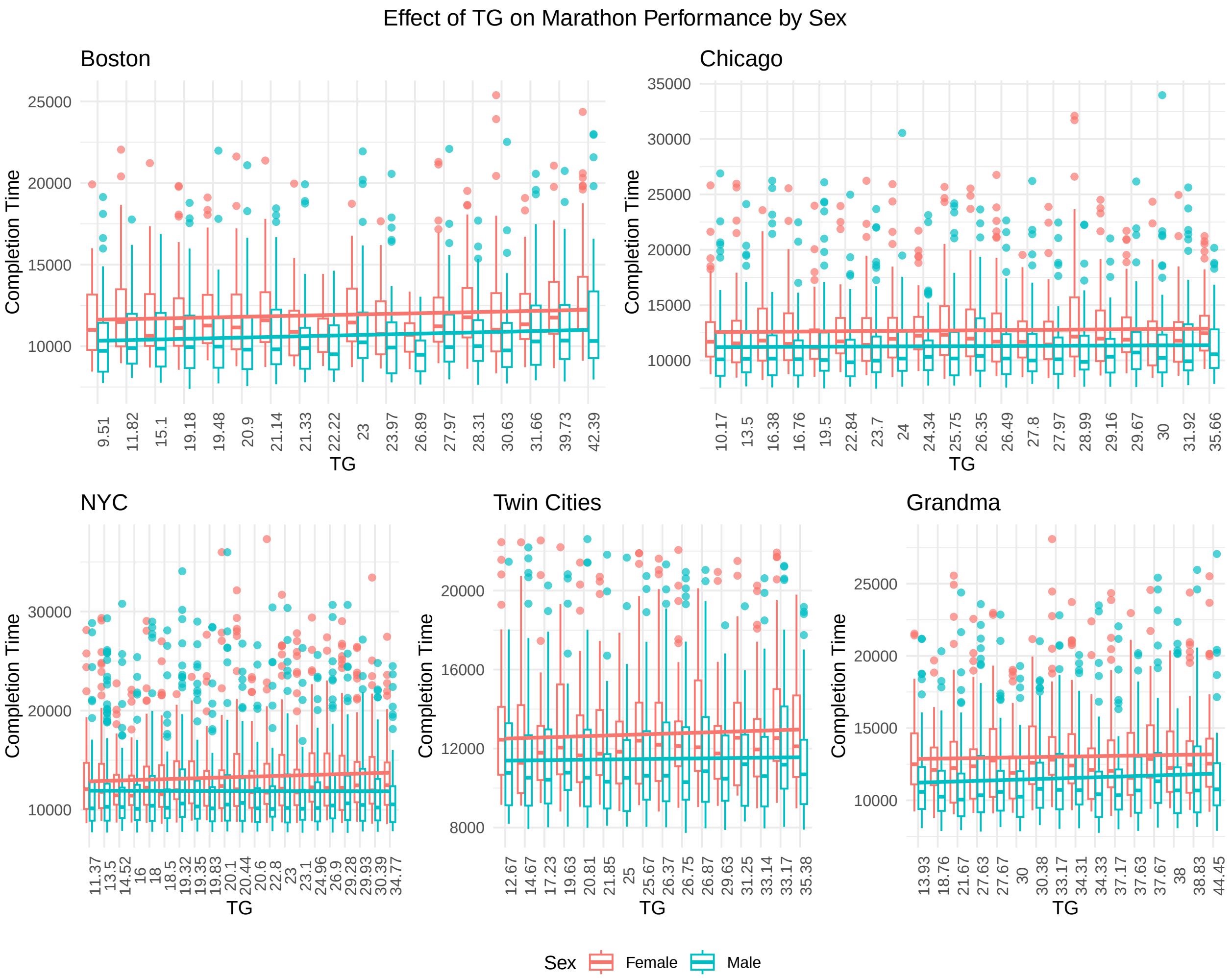
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



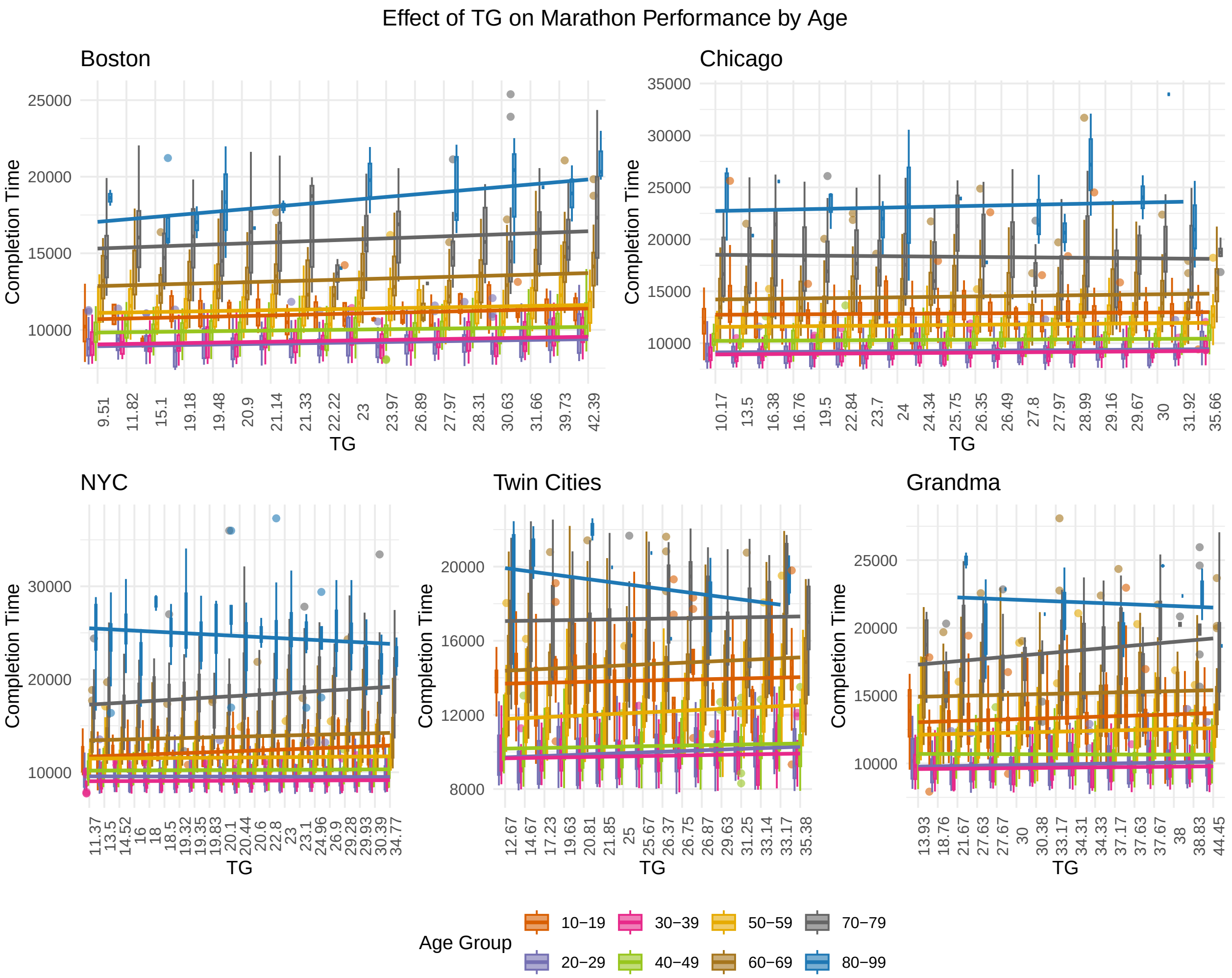
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



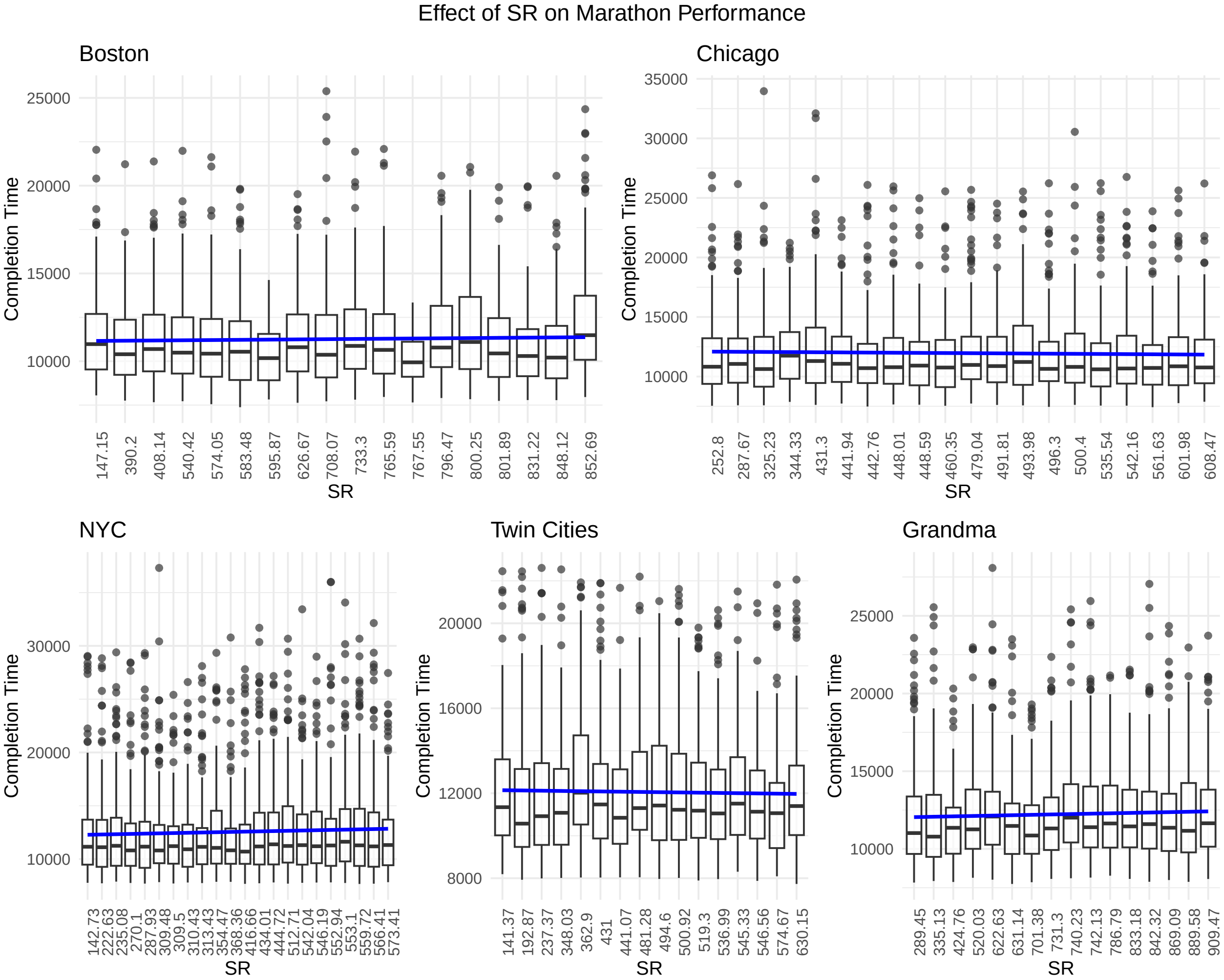
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



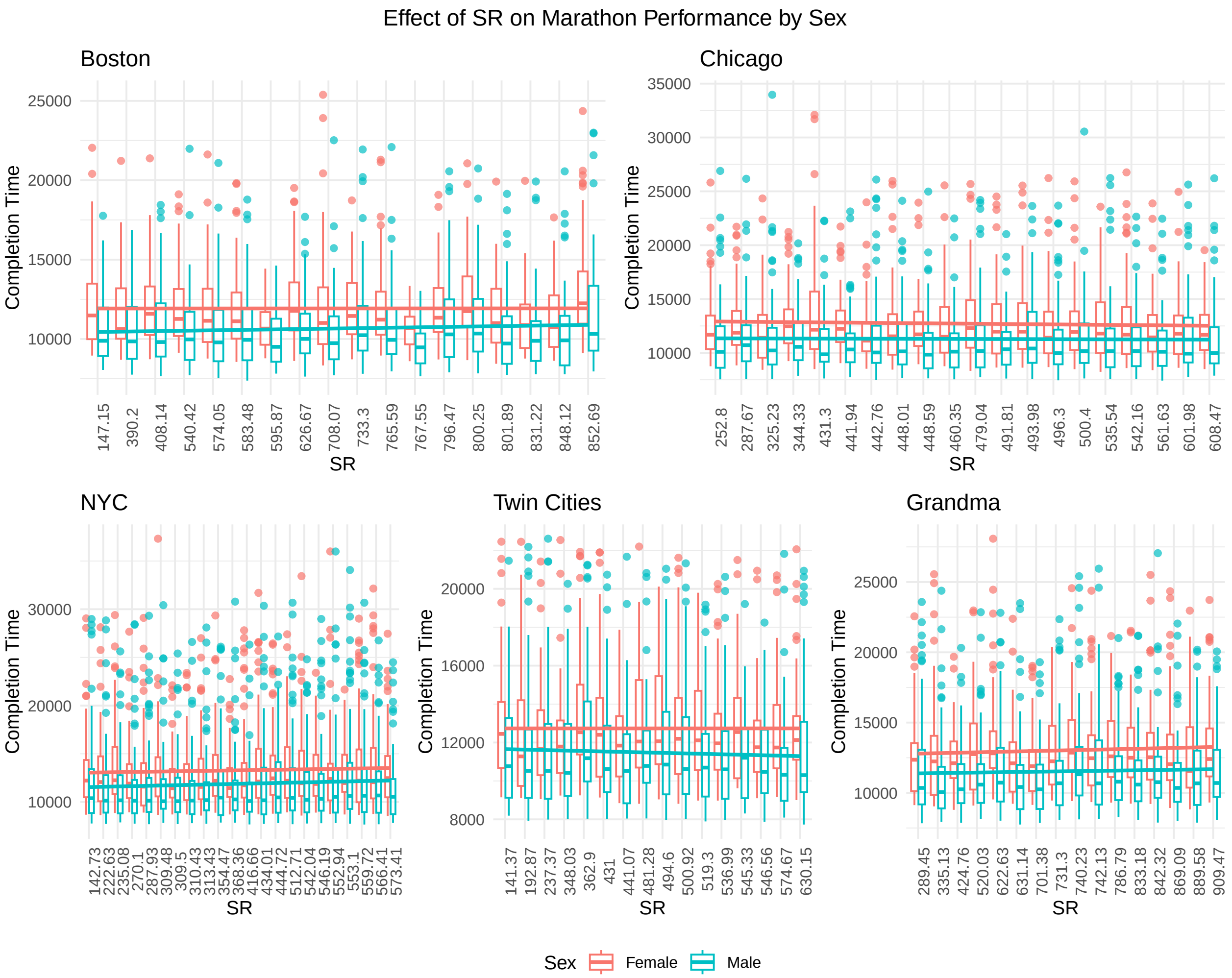
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



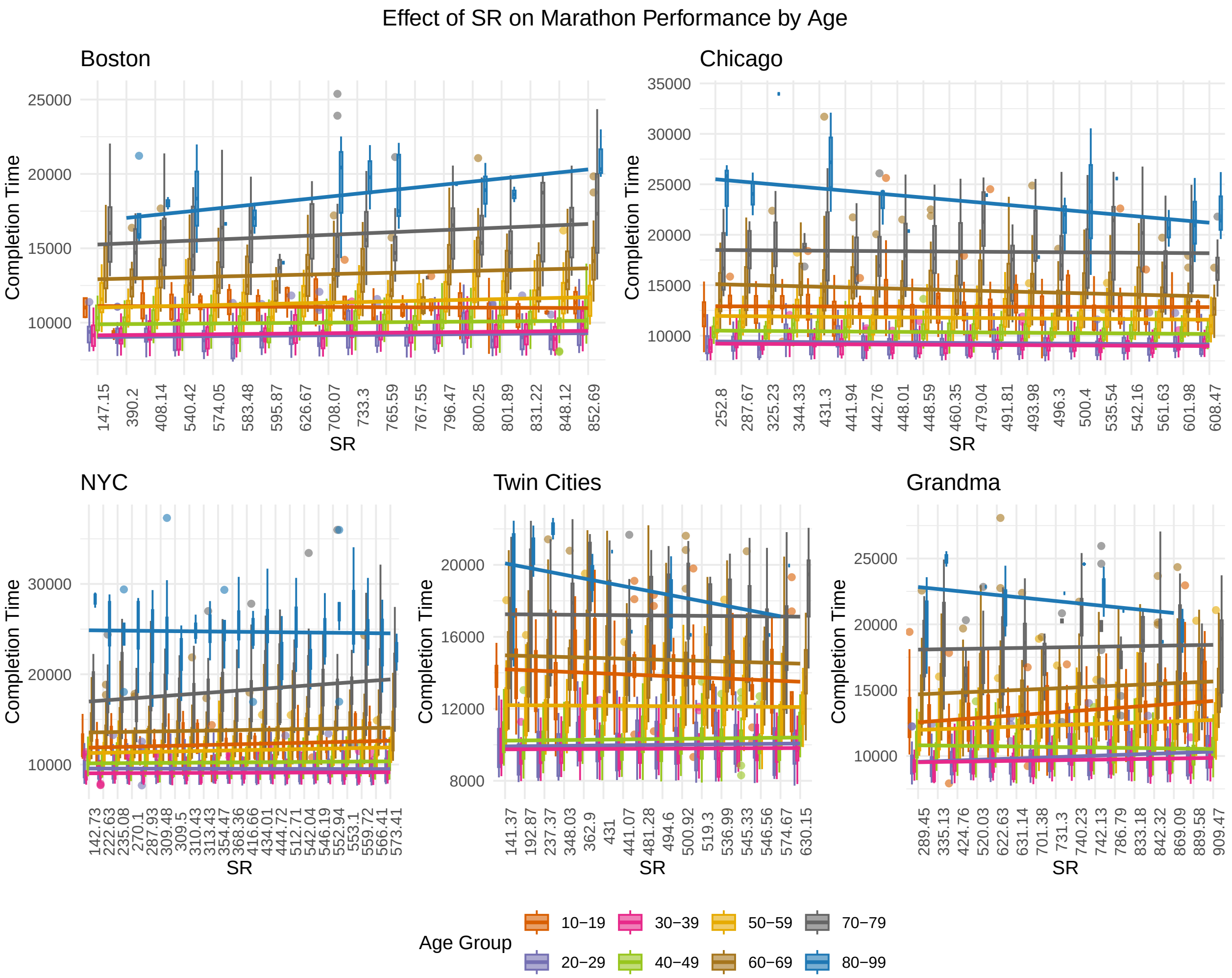
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



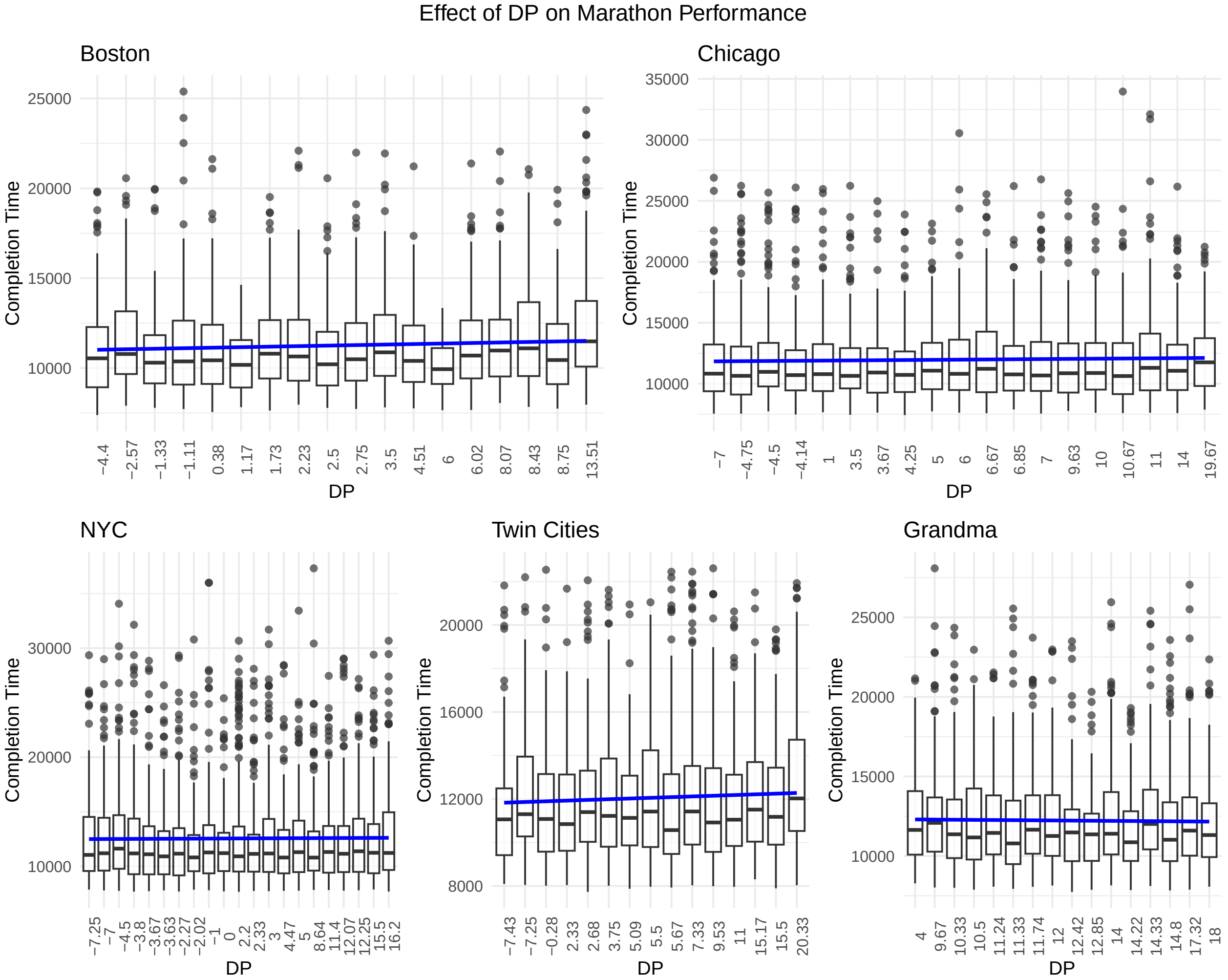
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



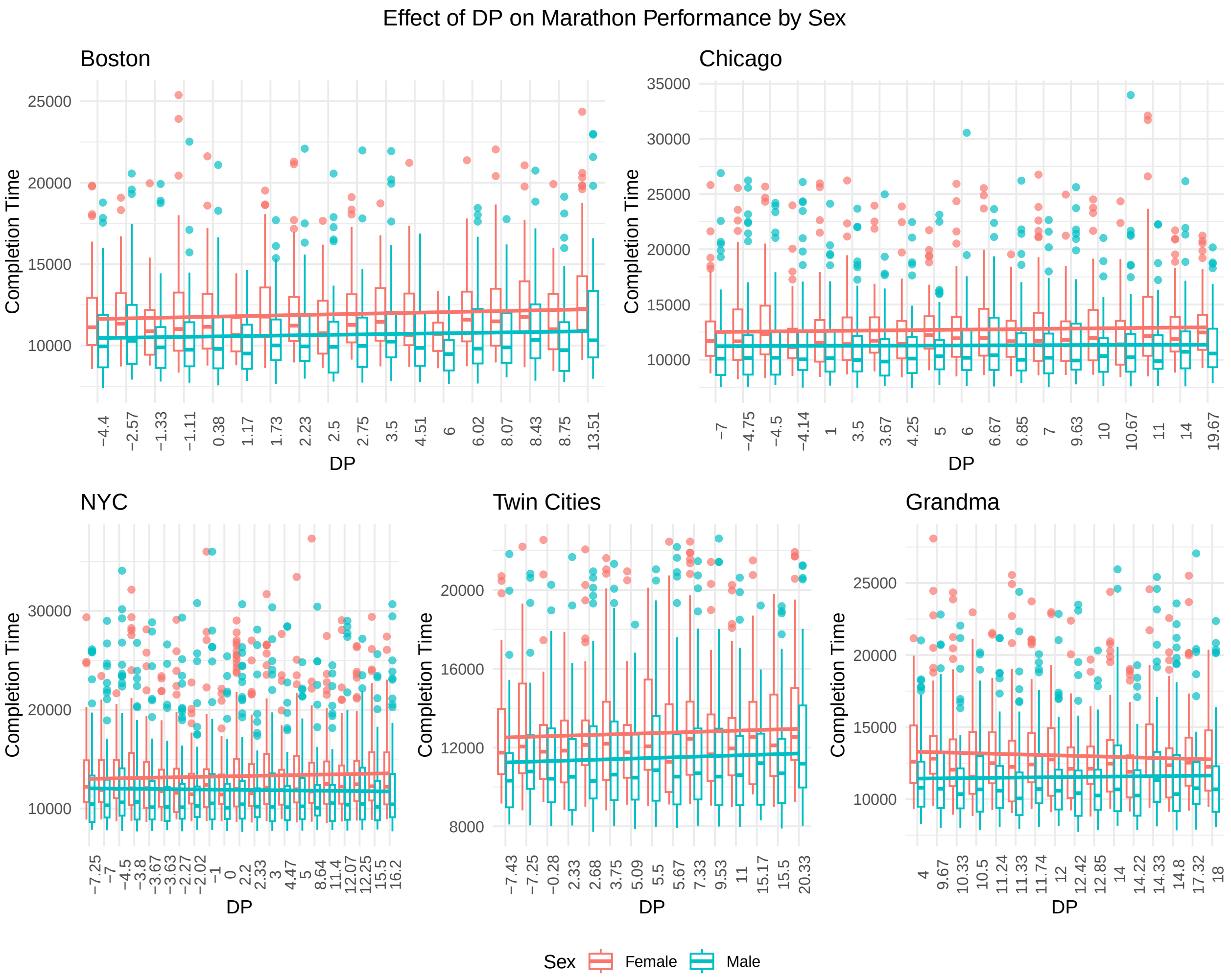
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



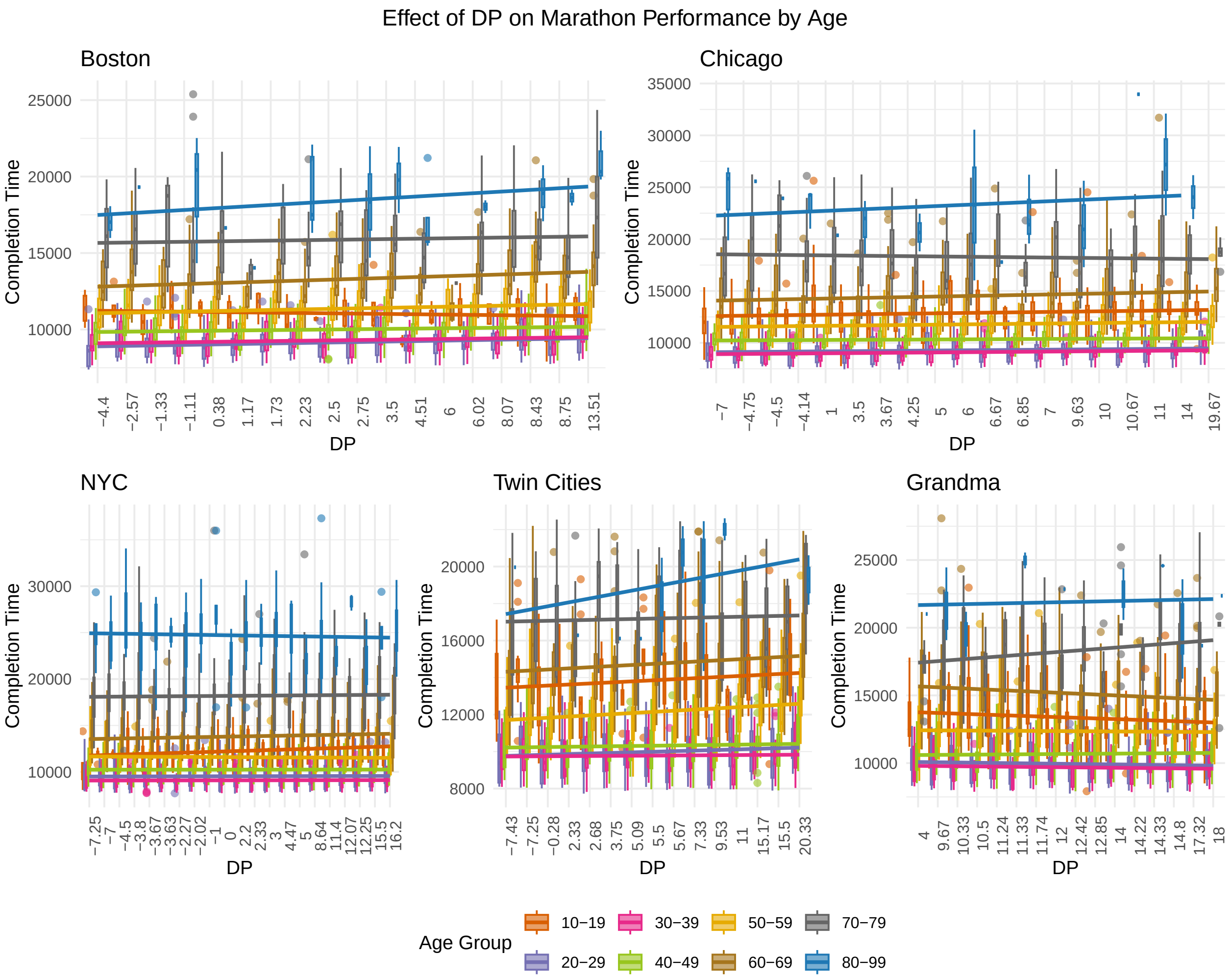
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



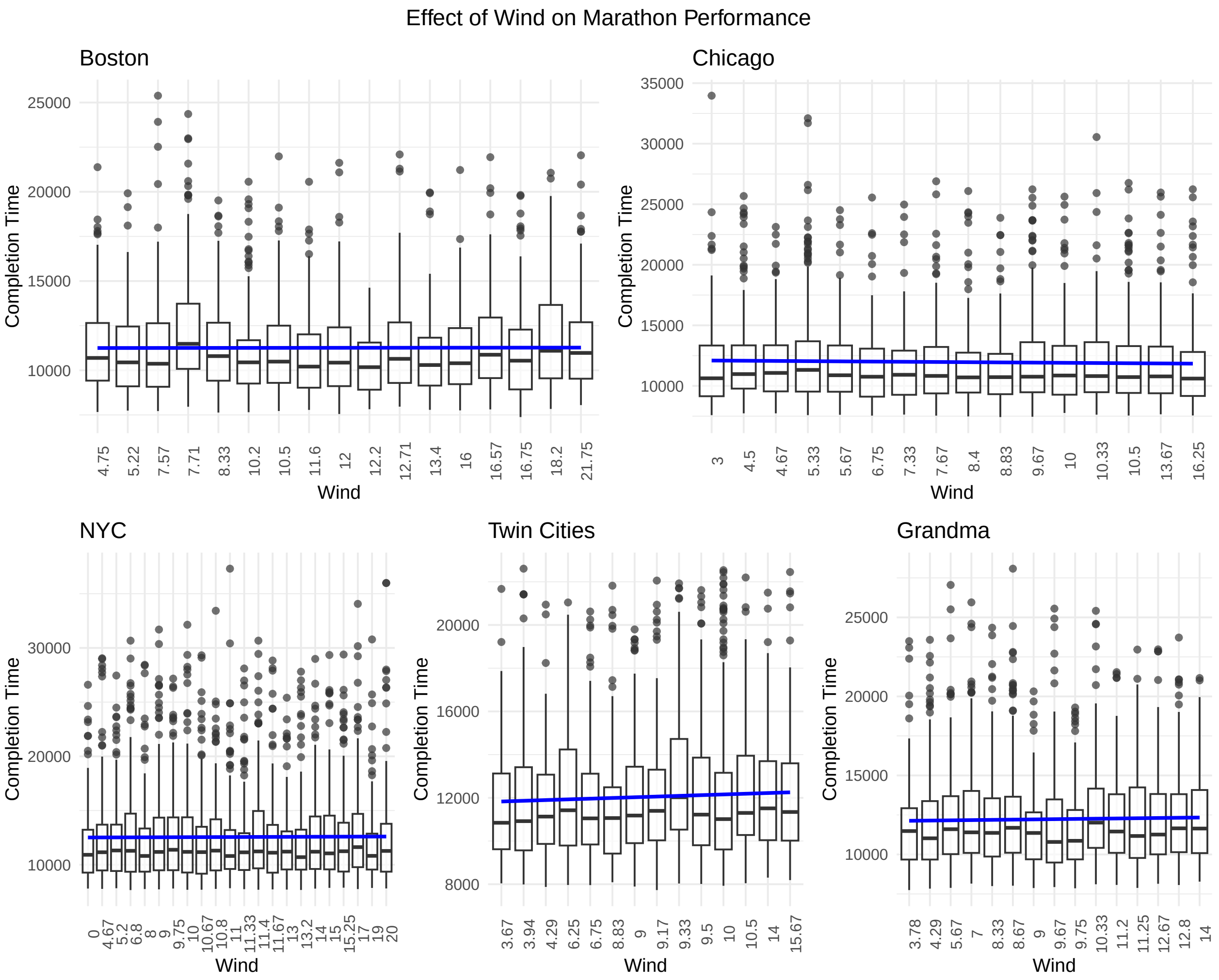
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



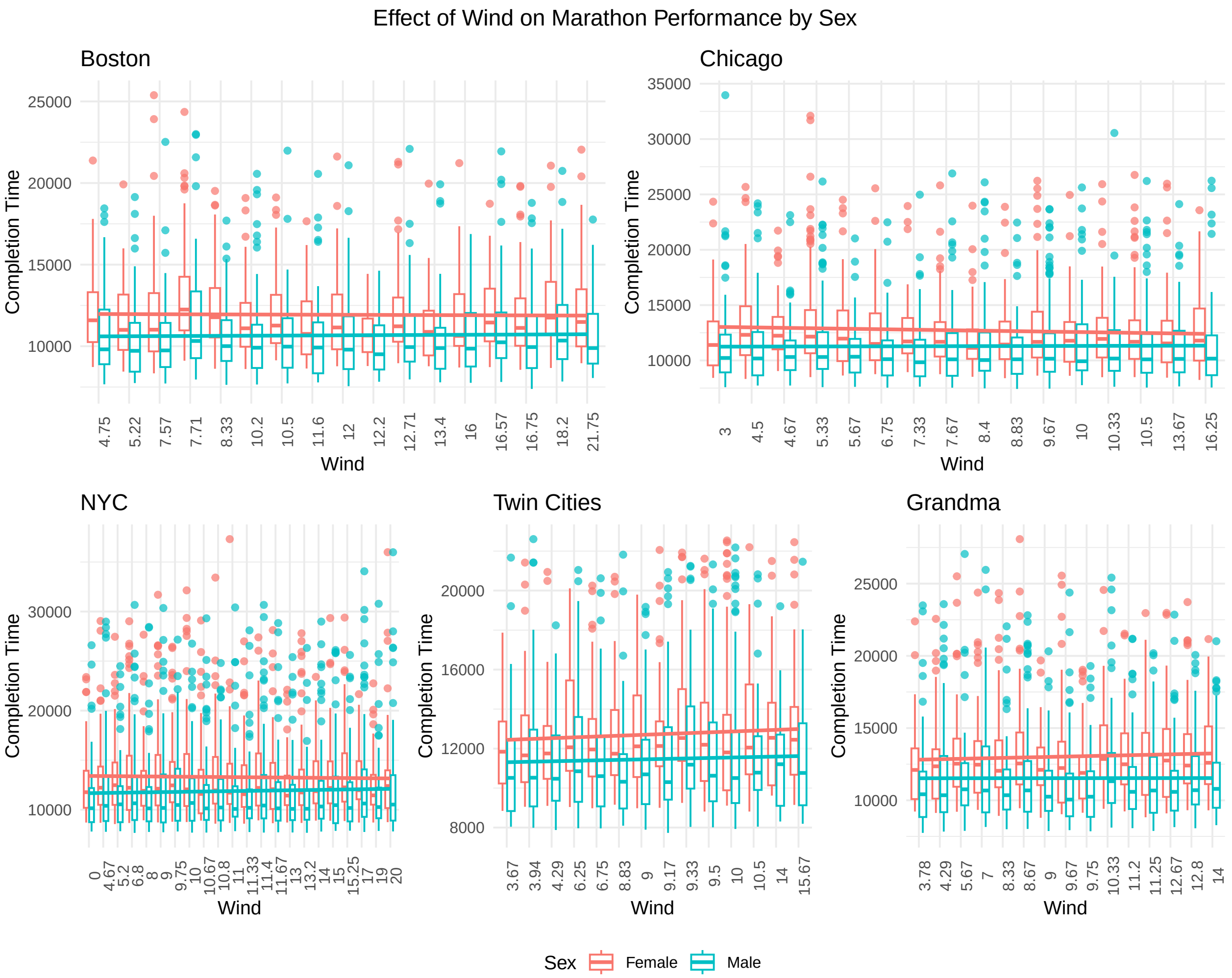
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



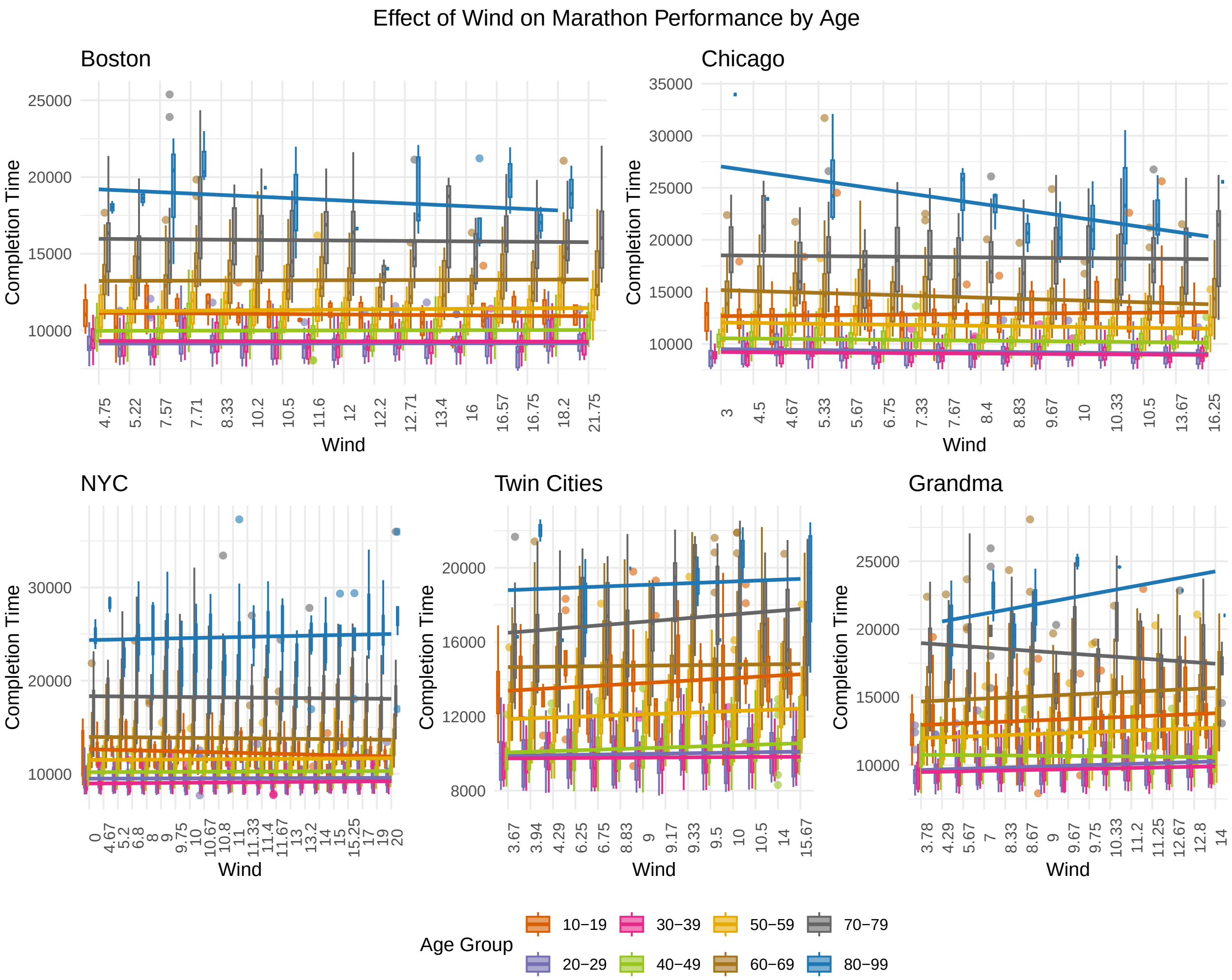
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



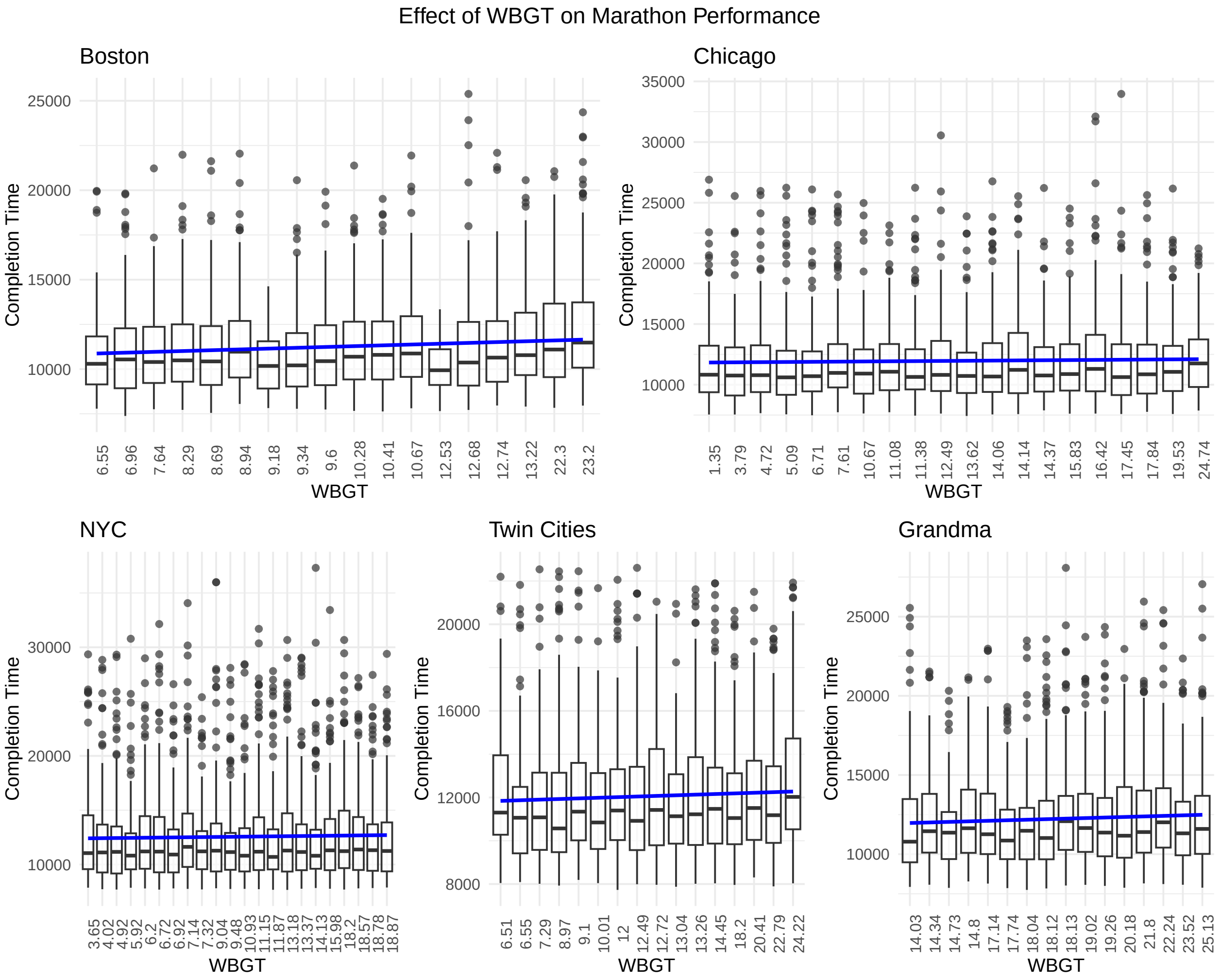
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



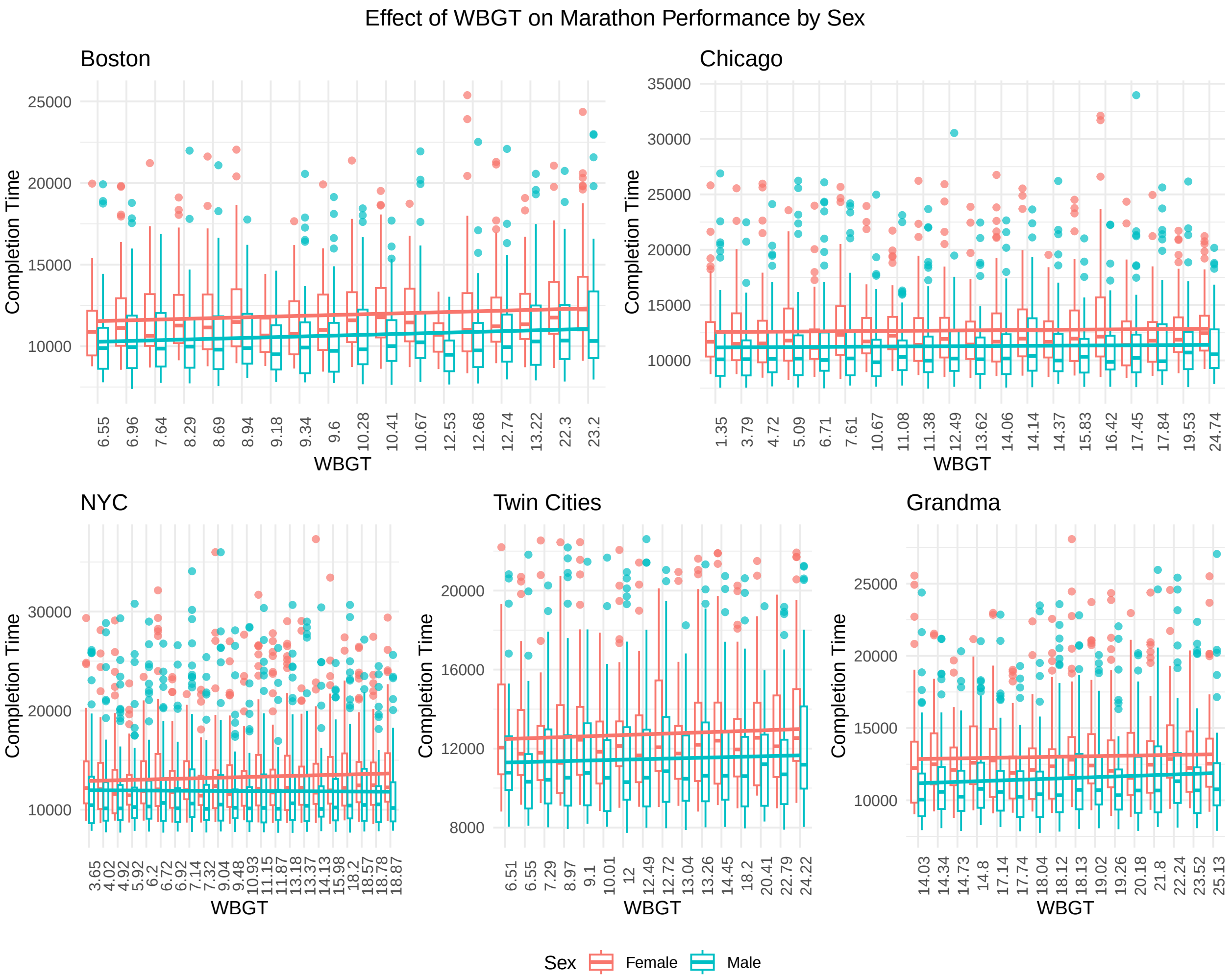
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



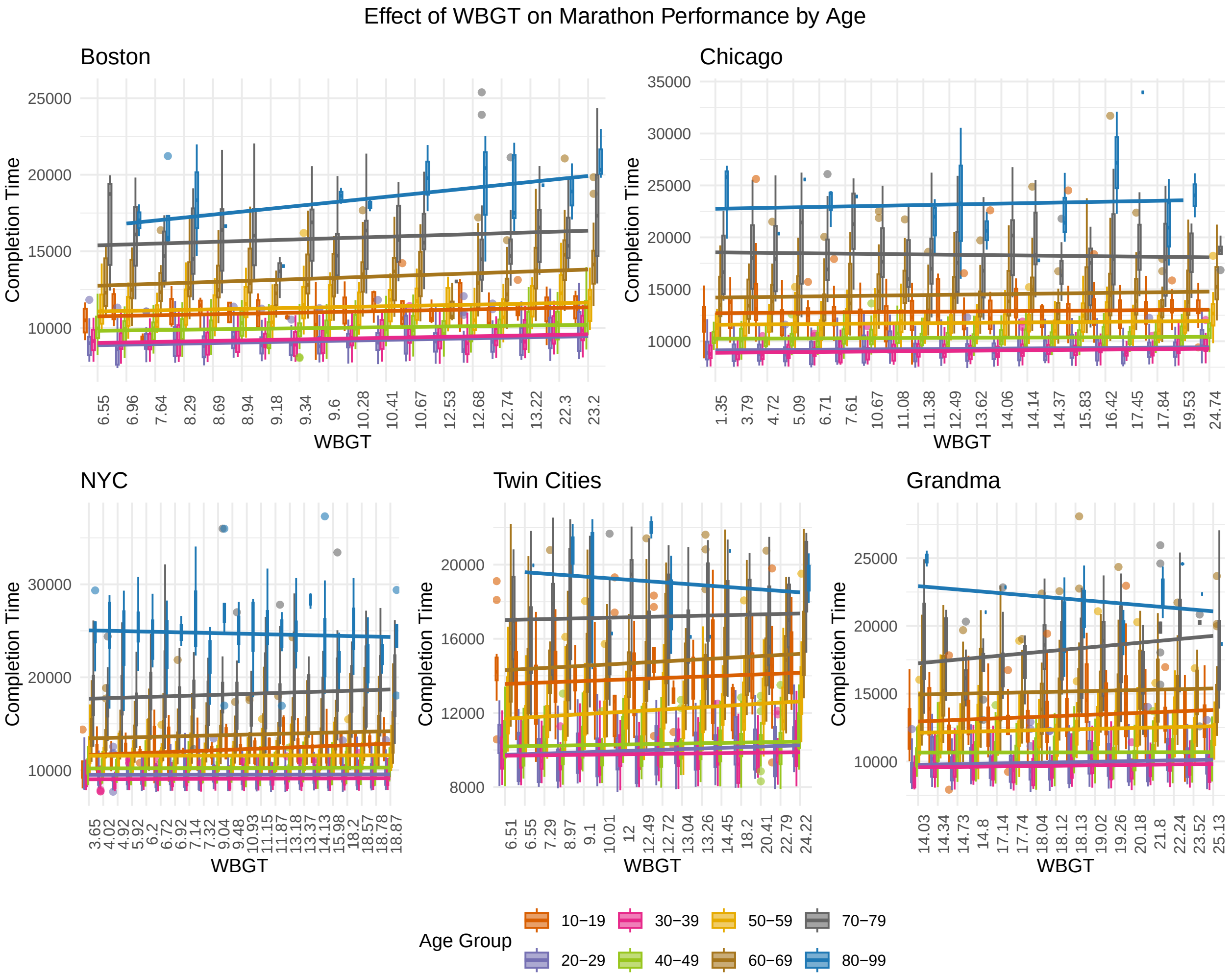
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



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)
library(patchwork)

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

```r
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)
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"
```

```r
# 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))
```

```r
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 = as.factor(Year), y = CR, fill = Sex)) +
  geom_boxplot() +
  facet_wrap(~ Race, scales = "free_y", nrow = 2) +
  labs(title = "Completetion Time Comparison by Sex",
       x = "Year",
       y = "CR",
       fill = "Sex") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1),
    legend.position = "top"
  )

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"
  )
weather_total_effects <- function(weather_var) {
  race_list <- c("Boston", "Chicago", "NYC", "Twin Cities", "Grandma")

  plot_list <- list()

  for (race in race_list) {
    plot <- marathon_data %>%
      filter(Race == race) %>%
      ggplot(aes(x = as.factor(round(!!sym(weather_var), 2)), y = CR)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = 1), method = "lm", color = "blue", se = FALSE) +
      labs(title = race) +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 90))

    plot_list[[race]] <- plot
  }

  combined_plot <- (plot_list$Boston | plot_list$Chicago) /
                   (plot_list$NYC | plot_list$`Twin Cities` | plot_list$Grandma) +
    plot_annotation(
      title = paste("Effect of", weather_var, "on Marathon Performance"),
      theme = theme(plot.title = element_text(hjust = 0.5))
    ) &
    labs(x = weather_var, y = "Completion Time")

  return(combined_plot)
}

weather_sex_effects <- function(weather_var) {

  race_list <- c("Boston", "Chicago", "NYC", "Twin Cities", "Grandma")

  plot_list <- list()

  for (race in race_list) {
    plot <- marathon_data %>%
      filter(Race == race) %>%
      ggplot(aes(x = as.factor(round(!!sym(weather_var), 2)), y = CR, color = Sex)) +
      geom_boxplot(alpha = 0.7) +
      geom_smooth(aes(group = Sex), method = "lm", se = FALSE) +
      labs(title = race,
           x = weather_var,
           y = "Completion Time") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

    plot_list[[race]] <- plot
  }

  combined_plot <- (plot_list$Boston | plot_list$Chicago) /
                   (plot_list$NYC | plot_list$`Twin Cities` | plot_list$Grandma) +
    plot_layout(guides = "collect", axis_titles = "collect") &
    theme(legend.position = 'bottom') &
    plot_annotation(
      title = paste("Effect of", weather_var, "on Marathon Performance by Sex"),
      theme = theme(plot.title = element_text(hjust = 0.5))
    ) &
    labs(x = weather_var, y = "Completion Time")

  return(combined_plot)
}
```

```r
weather_age_effects <- function(weather_var) {

  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")

  race_list <- c("Boston", "Chicago", "NYC", "Twin Cities", "Grandma")

  plot_list <- list()

  for (race in race_list) {
    plot <- marathon_data %>%
      filter(Race == race) %>%
      ggplot(aes(x = as.factor(round(!!sym(weather_var), 2)), y = CR, fill = Age_group, color = Age_grou
      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 = race,
           x = weather_var,
           y = "Completion Time (CR)",
           fill = "Age Group",
           color = "Age Group") +
      theme_minimal() +
      theme(
        axis.text.x = element_text(angle = 90)
      )

    plot_list[[race]] <- plot
  }

  combined_plot <- (plot_list$Boston | plot_list$Chicago) /
                   (plot_list$NYC | plot_list$`Twin Cities` | plot_list$Grandma) +
    plot_layout(guides = "collect", axis_titles = "collect") &
    theme(legend.position = 'bottom') &
    plot_annotation(
      title = paste("Effect of", weather_var, "on Marathon Performance by Age"),
      theme = theme(plot.title = element_text(hjust = 0.5))
    ) &
    labs(x = weather_var, y = "Completion Time")

  return(combined_plot)
}

weather_total_effects("TD")
weather_sex_effects("TD")
weather_age_effects("TD")
```

```
weather_total_effects("TW")
weather_sex_effects("TW")
weather_age_effects("TW")

weather_total_effects("RH")
weather_sex_effects("RH")
weather_age_effects("RH")

weather_total_effects("TG")
weather_sex_effects("TG")
weather_age_effects("TG")

weather_total_effects("SR")
weather_sex_effects("SR")
weather_age_effects("SR")

weather_total_effects("DP")
weather_sex_effects("DP")
weather_age_effects("DP")

weather_total_effects("Wind")
weather_sex_effects("Wind")
weather_age_effects("Wind")

weather_total_effects("WBGT")
weather_sex_effects("WBGT")
weather_age_effects("WBGT")
```