PHP 2550: Project 1

Peirong Hao

For this project, we collaborate with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. We aim to investigate the impact of age and weather, including temperature, humidity, solar radiation, and wind, on marathon performance in men and women. We have four datasets, including environmental conditions and participants' performances from five major marathons (Boston, Chicago, New York, Twin Cities, and Grandma's Marathons) from 1993 to 2016. In total, our data contain 11564 runners who participated in 98 marathons over the years. Regarding missingness, 16% of observations from the aqi dataset have missing values in the aqi column. Also, four matches (2011 Chicago, 2011 NYC, 2011 TC, 2012 Grandmas) have missing weather parameters. Since the missingness is unrelated to other data, we consider it missing completely at random, i.e., MCAR.

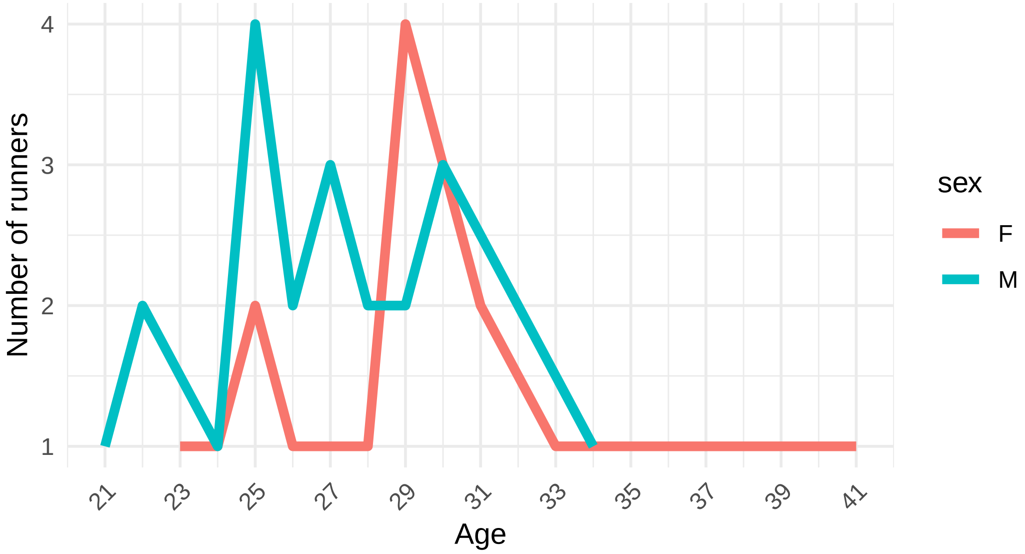

Figure 1: Number of people breaking marathon records by sex as age increases

For all five races, women aged between 18 and 60 participated in more races than the other age groups, and more men between 18 and 75 competed in races than the other age groups. Our first aim is to examine the effects of increasing age on marathon performance in men and women. CR is the course record for each year's marathon, obtained from last year's record. %CR refers to the percent course record. A positive %CR means the percentage more time used to complete the marathon. A negative %CR means the participant breaks the record. We can calculate the completion time for each individual using CR * (1+%CR). Figure 1 displays the number of people who broke marathon records. The record-breakers were between the ages of 21 and 41. Males broke the record in a younger age range, but females could break the record in the broader age group.

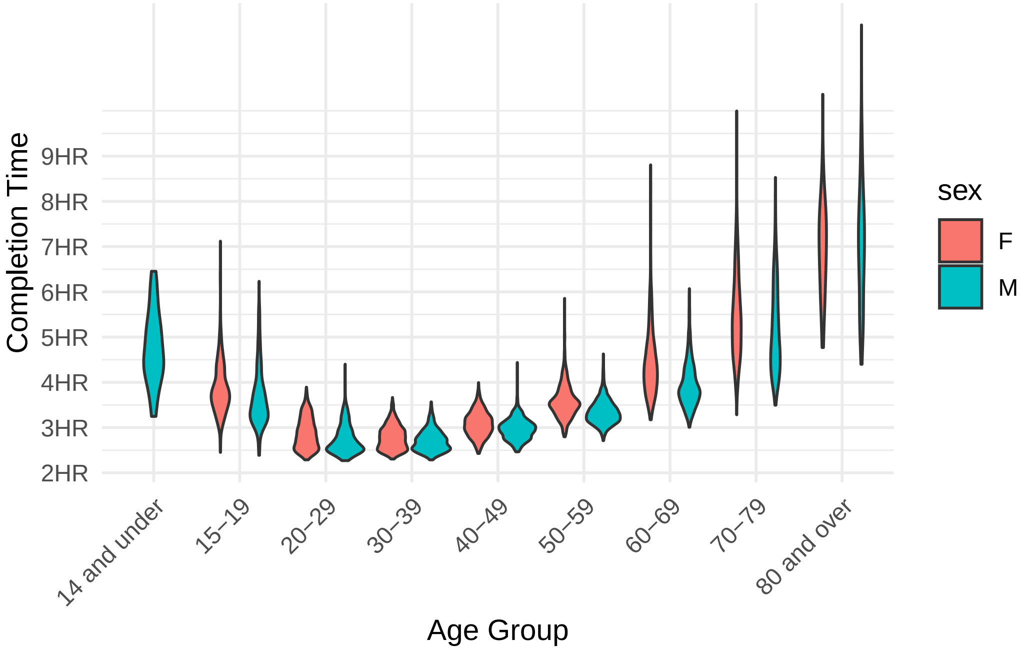

Figure 2: Distribution of completion time for each age group

Figure 2 shows a corresponding trend: people between 20 and 40 ran much faster than the other age groups. Regarding the completion time of each match, some people in the elder age group took longer than 9 hours to complete, with the longest time being 11 hours and 54 minutes. These outliers could be due to some data collection error since Boston, Twin Cities, and Chicago marathons must be completed within 6 hours, 6 hours, and 15 minutes, 6 and a half hours, respectively. However, Grandma's and NYC marathons allow participants to continue finishing on the sidewalk after 7 hours and 8 hours limit, respectively, without services, including cross-street protection, medical assistance, and aid stations. Thus, even though those outliers weren't official finishing times according to very well fit, they could be valid self-reported completion times.

Our second aim is to explore the impact of environmental conditions on marathon performance and examine if the effect differs across age and gender. We focus on a few weather parameters. First of all, Wet Bulb Globe Temperature (WBGT) is an informative measurement since it is derived from three different sources: Wet bulb temperature (Tw, which accounts for humidity), Black globe temperature (Tg, which accounts for solar radiation), and Dry bub temperature (Td). Flag, which is determined by WBGT and the chance of heat illness, is another option. The Flag variable has five categories arranged from best to worst condition: White, Green, Yellow, Red, and Black. We do not have any data falling in the Black condition, probably because Marathons are canceled in extreme weather.

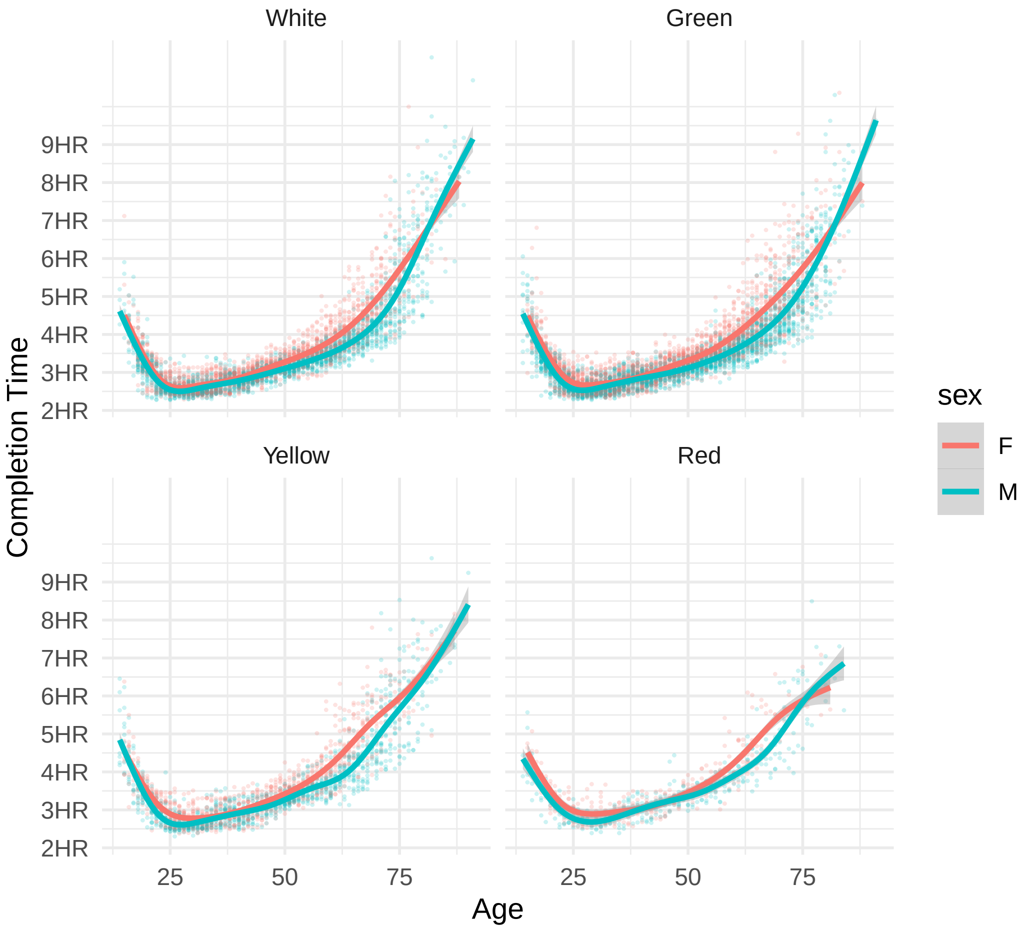

Figure 3: Impact of flag on completion time as age increases

Figure 3 demonstrates that completion time increases as age increases, and there is little variation in the trend among the various Flag groups. Before the age of 50, there is no remarkable difference in the completion times of men and women. Women took slightly longer to finish between 50 and 75, but the pattern flips around the age of 77.

According to Figure 4, more people took part when the Flag was green, and a far lower number did so when it was red. Although the number of participants was significantly lower when the Flag was yellow or red, those who did participate may have been more motivated and had more training than those who did not, allowing them to finish the marathons in a comparable

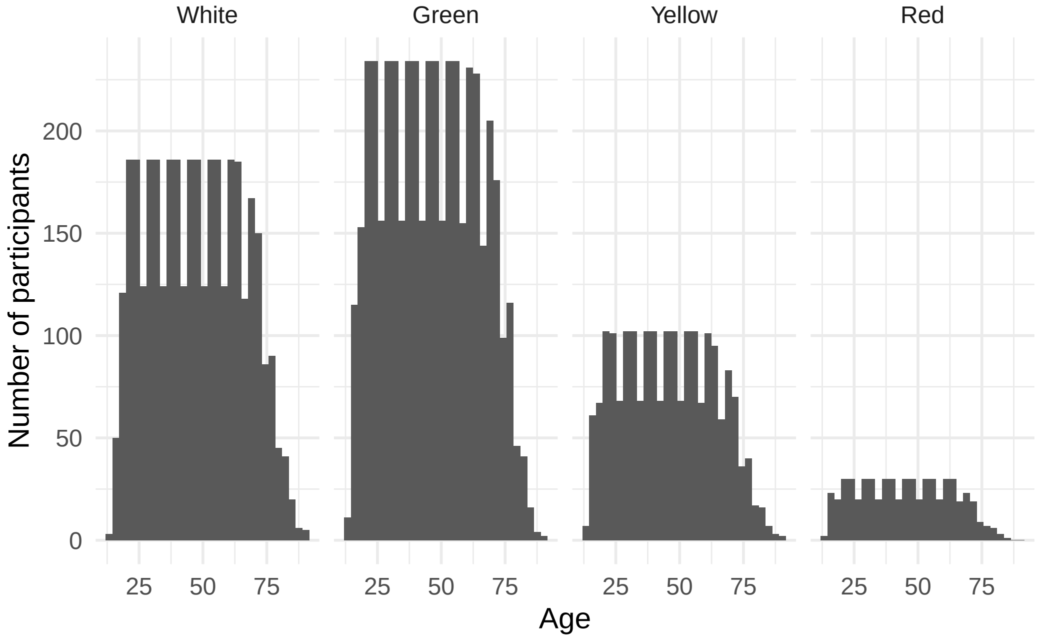

Figure 4: Impact of flag on participation as age increases

time frame.

Another important weather parameter is wind speed, measured in km/h. There is a slight difference between the effects of wind speed on completion time for men and women. Based on Figure 5, NYC participants took roughly 30 minutes longer to complete tasks than Boston participants on average. This might result from the events being hosted at different times of the year: Boston in April, Grandmas (in Minnesota) in June, Twin Cities and Chicago in October, and New York City in November. All of the locations above are in the northern portion of the United States; therefore, in the winter, the wind and snowy roads of New York City affected runners more than the bright, breezy summers of Boston.

The air quality index (aqi), which is present in the data, has a lot of missing values. Hence, we choose another variable called arithmetic_mean, which is used to calculate aqi instead. Ozone and PM2.5 measurements are associated with parameter codes 44201 and 88101. It is important to note that although PM2.5 is coded in 88502, it has many more missing values than 88101, the primary PM2.5 measurement. Because Ozone and PM2.5 measurements were recorded at different sublocations and times of the day for each city, we calculate an average of each parameter for each marathon date and site.

Ozone Solutions suggested the safe limit for ozone is 0.1 ppm, and our Ozone data were well below that threshold. The relationships between Completion Time and Ozone held constant for most age groups but varied slightly for the youngest and eldest groups in Figure 6. According to Indoor Air Hygiene Institute, PM2.5 at or under 12 micrograms/cubic meter is considered

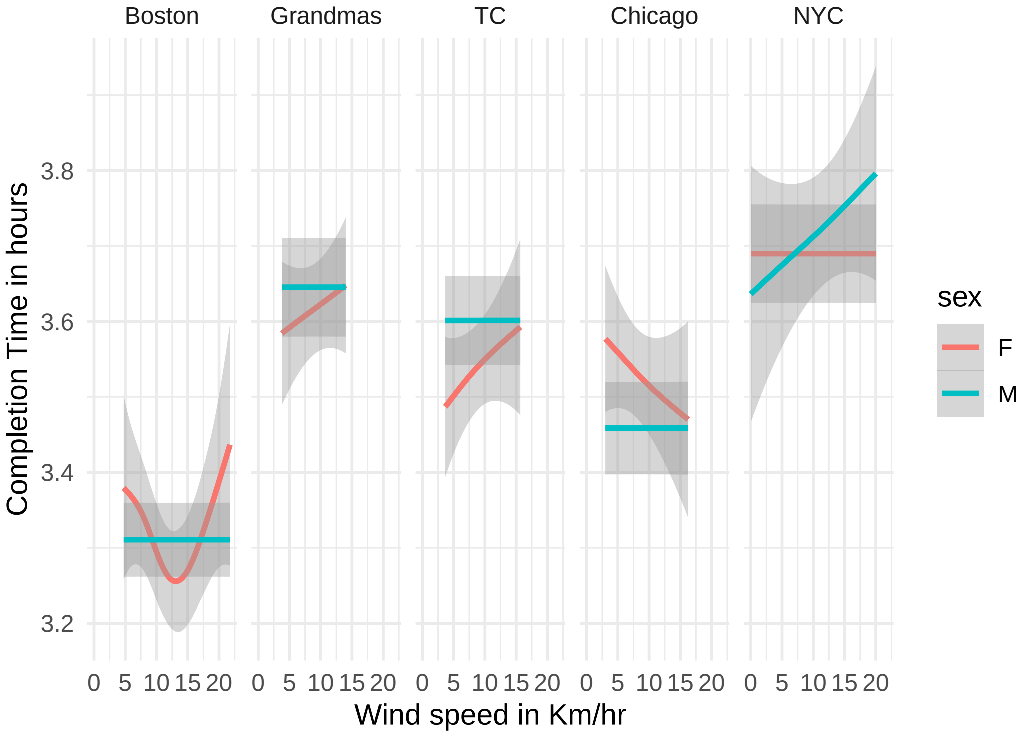

Figure 5: Impact of wind on completion time

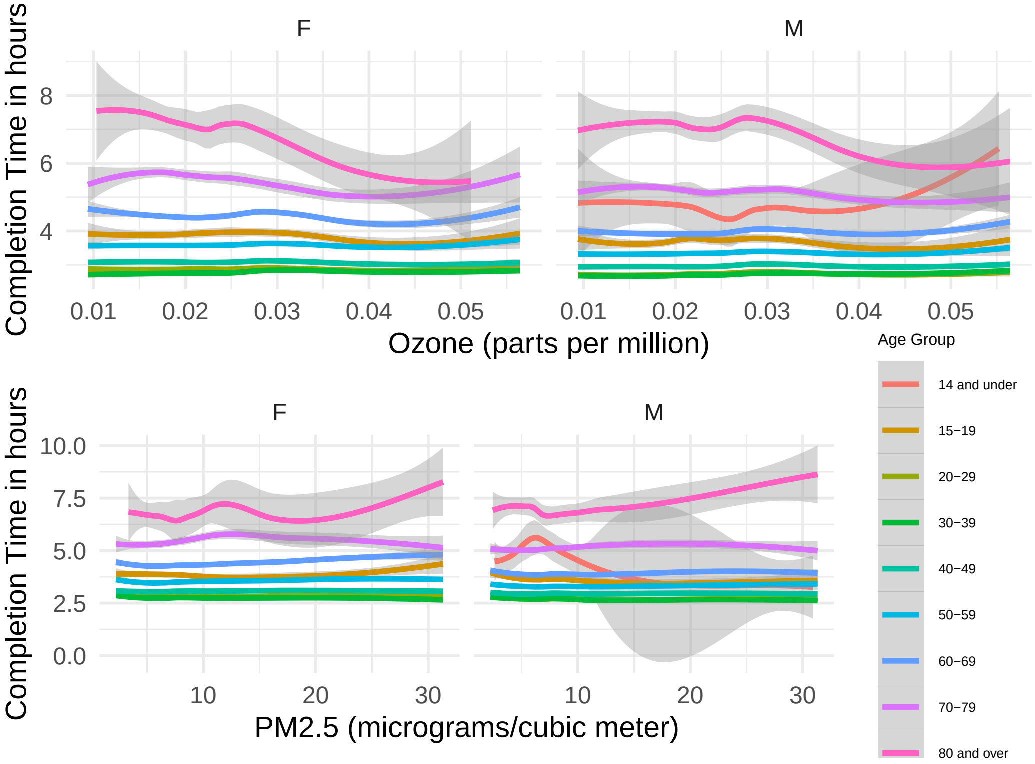

Figure 6: Impact of air pollutants' condition on completion time

healthy, but it is deemed unhealthful when it reaches or surpasses 35 micrograms/cubic meter. Because the recorded PM2.5 data were below 35 micrograms/cubic, there was not much health concern about it. Concerning variation in completion time with PM2.5, there was little change for the majority but slight variation for the youngest and eldest groups. These relations held for both sex groups.

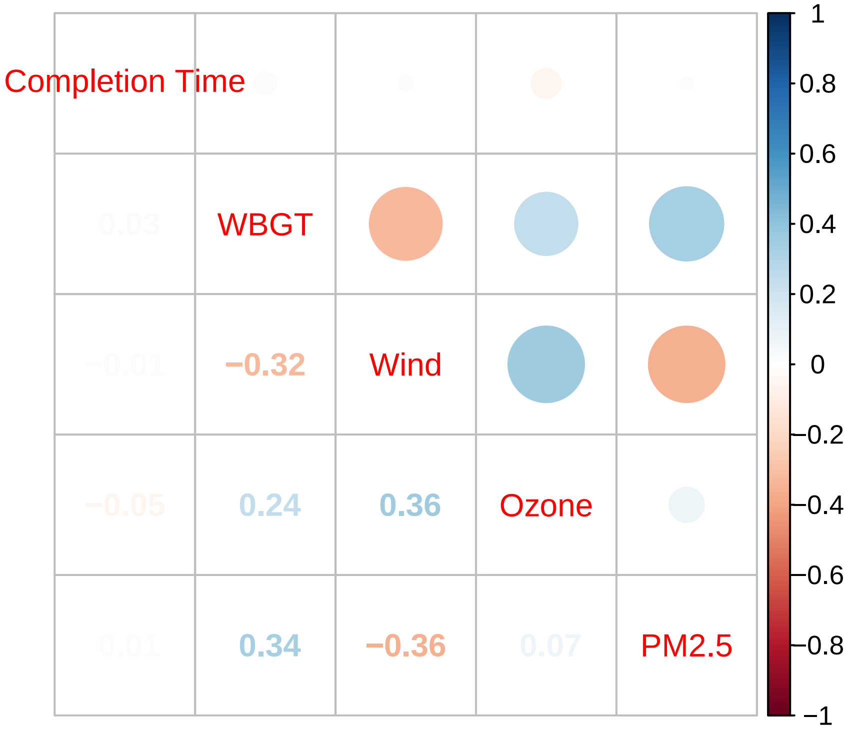

Figure 7: Correlation plot

Our last aim is to identify the weather parameters most affecting marathon performance. There is no weather parameter that has a strong correlation with Completion Time, and there is no very strong correlation between weather parameters either. In this case, we have Completion Time (Y) and try to find its linear relationship with other quantitative weather variables, age, and sex (X). Best subset, forward, and backward selection all result in the model: Completion Time = 5537.299 + 73.790 * WBGT + 31.721 * Wind - 30594.337 * Ozone - 403.091 * sexM + 150.287 * age.

Table 1: Coefficients of linear model

	Coefficients
(Intercept)	5537.299
WBGT	73.790
Wind	31.721

	Coefficients
Ozone	-30594.337
sexM	-403.091
age	150.287

The interpretation of the model is as follows: for each additional unit increase in WBGT, the completion time is expected to increase by 73.790 seconds, holding all other variables constant; for each additional unit increase in wind speed, the completion time is expected to increase by 31.721 seconds, holding all other variables constant; for each additional unit increase in ozone, the completion time is expected to decrease by 30594.337 seconds, holding all other variables constant; for each additional unit increase in age, the completion time is expected to increase by 150.287 seconds, holding all other variables constant; for a male runner, we expect completion time to decrease by 403.091 seconds compared to a female, holding other variables constant.

For this model, the overall F-test statistic is 2176, which is associated with a p-value <2.2e-16. Since the p-value $<\alpha$=0.05, we reject the null hypothesis and conclude that this model is in fact significant. There is a relationship between Y (completion time) and at least some of the predictor variables (WBGT, Wind, Ozone, Sex, Age). When looking at the t-tests on the slopes for predictor variables, they have small p-values ($<\alpha$=0.05), thus they are significant predictors. Coefficient of determination is the Adjusted R-squared value, 49.55% of the variability in completion time (Y) is explained by its linear relationship with WBGT, Wind, Ozone, Sex, and Age (X).

Conclusion and Limitations

People aged between 20 and 40 run much faster than the other age groups and could break course records. Age, wind speed, and WBGT all have a positive linear relationship with completion time: older runners finish the marathon later; severe weather, such as strong winds and higher temperatures, also delays completion. Additionally, we discovered that, on average, men run faster than women. One strange finding is that as ozone levels rise, individuals finish marathons faster. This could result from the observed ozone level falling between 0 and 0.054 ppm, which is thought to be good air quality (Kaiterra). Furthermore, PM2.5 is not considered a significant predictor of completion time. Our linear model's modified R-squared of 49.55% is relatively low. Four marathons had incomplete weather data. If we conduct imputation for missing values, we can get more robust results. It is also worth examining more marathons in different regions to investigate the impact of weather on performance further.

Code Appendix

```r
set.seed(123456)
library(tidyverse)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(readr)
library(visdat)
#library(naniar)
#library(patchwork)
library(gridExtra)
library(corrplot)
library(olsrr)# model selection

#read in data files
df.aqi<-read.csv("aqi_values.csv")
df.course.record<-read.csv("course_record.csv")
df.marathon.dates<-read.csv("marathon_dates.csv")
df.project1.main<-read.csv("project1.csv")
df.aqi<-df.aqi %>%
  mutate(race = case_when(
          marathon == "Boston" ~ "Boston",
          marathon == "Chicago" ~ "Chicago",
          marathon == "Grandmas" ~ "Grandmas",
          marathon == "NYC" ~ "NYC",
          marathon == "Twin Cities" ~ "TC",
          .default = NA)) %>%
  select(-c(marathon))

df.course.record<-df.course.record %>%
  mutate(race = case_when(
          Race == "B" ~ "Boston",
          Race == "C" ~ "Chicago",
          Race == "D" ~ "Grandmas",
          Race == "NY" ~ "NYC",
          Race == "TC" ~ "TC",
          .default = NA)) %>%
  select(-c(Race))
```

```
df.marathon.dates<-df.marathon.dates %>%
  mutate(race = case_when(
          marathon == "Boston" ~ "Boston",
          marathon == "Chicago" ~ "Chicago",
          marathon == "Grandmas" ~ "Grandmas",
          marathon == "NYC" ~ "NYC",
          marathon == "Twin Cities" ~ "TC",
          .default = NA)) %>%
  select(-c(marathon))

df.project1.main<-df.project1.main %>%
  mutate(race = case_when(
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 0 ~ "Boston",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 1 ~ "Chicago",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 4 ~
 ↪  "Grandmas",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 2 ~ "NYC",
          Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D. == 3 ~ "TC",
          .default = NA)) %>%
  select(-c(Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.))
# Visualize missing data pattern for all dfs

#aqi column has 16% missingness
df.aqi %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

#no missingness
df.course.record %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

#no missingness
df.marathon.dates %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")

#last 8 columns have missingness
df.project1.main %>% abbreviate_vars() %>% vis_miss() + theme(text =
 ↪  element_text(size = 7)) + ggtitle("Missing Data")
miss_tbl<-df.project1.main %>% filter(if_any(everything(), is.na))%>%
  mutate(race = as.factor(race))%>%
  group_by(Year, race)%>%
  summarise(n = n(),.groups = 'drop')%>%
  rename(Race = race, `Number of observations` = n)
kable(miss_tbl, booktabs = TRUE, digits = 2)
#MCAR
#start to join datasets
df.all <- left_join(df.project1.main, df.course.record,
 ↪  by=c("Year","race"))%>%
  arrange(Year, race)

#rename sex, age column
df.all<-df.all %>%
  mutate(sex = case_when(
          Sex..0.F..1.M. == 0 ~ "F",
          Sex..0.F..1.M. == 1 ~ "M",
          .default = NA))  %>%
  select(-c(Sex..0.F..1.M.))  %>%
  rename(age = Age..yr.)

trend.record.break <-df.all %>%
  filter(X.CR < 0) %>%
  select(sex, X.CR, age) %>%
  group_by(age, sex) %>%
  summarise(n = n(),.groups = 'drop') %>%
  arrange(sex, age) %>%
  mutate(sex = factor(sex))

# Trend plot - Break record
ggplot(trend.record.break)+
  geom_line(aes(x=age, y = n, color=sex, group=sex), linewidth=1.7)+
    scale_x_continuous(breaks = seq(21, 41, by = 2), guide =
     ↪  guide_axis(angle = 45))+
  theme_minimal()+
  ylab("Number of runners")+
  xlab("Age")
#actual time = CR + %CR * CR -> marathon performance
#CR, unique value for each race, each year
#convert CR from character to time

#https://www.runnersworld.com/races-places/a20794726/age-and-weight-groups-com/
#create age divisions
df.all<-df.all %>%
  mutate(CR = as.numeric(hms(CR)),
    actual.time = CR * (1+ X.CR/100),
    age.division = case_when(
          age <= 14 ~ "14 and under",
          age >= 15 & age <= 19 ~ "15-19",
          age >= 20 & age <= 29 ~ "20-29",
          age >= 30 & age <= 39 ~ "30-39",
          age >= 40 & age <= 49 ~ "40-49",
          age >= 50 & age <= 59 ~ "50-59",
          age >= 60 & age <= 69 ~ "60-69",
          age >= 70 & age <= 79 ~ "70-79",
          age >= 80 ~ "80 and over",
          .default = NA),
        age.division =  factor(age.division, levels = c("14 and under",
        "15-19","20-29","30-39","40-49","50-59","60-69","70-79","80 and
         ↪  over")),
        Flag = factor(Flag, levels = c("White", "Green", "Yellow",
         ↪  "Red", "Black")))

ggplot(df.all, aes(x=age.division, y=actual.time, fill=sex)) +
  geom_violin() +
  scale_y_continuous(breaks = c(3600*c(2,3,4,5,6,7,8,9)),
                     labels =
                      ↪  c("2HR","3HR","4HR","5HR","6HR","7HR","8HR","9HR"))+
  theme_minimal()+
  labs(x = "Age Group", y = "Completion Time")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
tbl.time<-t(summary(as.period(df.all$actual.time)))
kable(tbl.time, booktabs = TRUE, digits=2)
# summary table - participants overall
# Select relevant columns, Create Summary Table
trend.age.impact.summ<-df.all %>%
  select(sex, actual.time, age) %>%
  group_by(age, sex) %>%
  summarise(n = n(),.groups = 'drop') %>%
  # arrange(sex, age) %>%
  mutate(sex = factor(sex))

trend.age.impact.summ
#age: 14-91

# Trend plot - participants overall
```

```
ggplot(trend.age.impact.summ)+
  geom_line(aes(x=age, y = n, color=sex, group=sex), linewidth=1.7)+
    scale_x_continuous(breaks = seq(14, 92, by = 2), guide =
     ↪  guide_axis(angle = 45))+
  ylab("Number of participants")+
  xlab("Age")

# Trend plot - participants by region
df.all %>%
  ggplot(aes(x = age, colour = sex)) +
  facet_wrap(vars(race), ncol = 3) +
  geom_point(stat = "count",size=1) +
  geom_line(stat = "count") +
  labs(x = "Age", y = "Number of participants")
#aqi dataset, parameter_code column: 44201: Ozone, 88101 (primary
 ↪  measure)/88502: PM2.5, multiple rows correspond to one observation
#aqi: air quality index, calculated from mean (integer)
#arithmetic_mean: actual measurement, average sample across that time
 ↪  period, use this instead of aqi; units_of_measure
#do not need "sample_duration"
#WBGT, Flag = 0.7 * Tw + 0.2 * Tg + 0.1 * Td

#merge the other two dfs
df.all <- left_join(df.all, df.marathon.dates,
 ↪  by=c('Year'='year',"race"))%>%
  arrange(Year, race)

df.aqi$date_local<-as.Date(df.aqi$date_local)
df.all$date<-as.Date(df.all$date)
date.unique.list<-unique(df.all$date)

#pivot wider df.aqi, take mean of arithmetic mean for each region
df.aqi.wider <- df.aqi %>%
  filter(date_local %in% date.unique.list) %>%
  #select(date_local, parameter_code, arithmetic_mean, race,
   ↪  units_of_measure) %>%
  group_by(date_local, race, parameter_code) %>%
  summarise(mean=mean(arithmetic_mean),.groups = 'drop') %>%
  pivot_wider(names_from = parameter_code, values_from = mean)
```

```
df.all <- left_join(df.all, df.aqi.wider,
 ↪  by=c('date'='date_local',"race"))%>%
  arrange(Year, race)
ggplot(data = na.omit(df.all[,c("Flag","actual.time","sex","age")]),
       aes(x = age, y = actual.time, color = sex)) +
  geom_point(size=0.1,alpha=0.2) +
  stat_smooth() +
  facet_wrap(~Flag)+
  theme_minimal()+
  scale_y_continuous(breaks = c(3600*c(2,3,4,5,6,7,8,9)),
                     labels =
                      ↪  c("2HR","3HR","4HR","5HR","6HR","7HR","8HR","9HR"))+
  labs(x = "Age", y = "Completion Time")
#summary table
# df.all %>%
#   select(age.division, Flag) %>%
#   group_by(Flag, age.division) %>%
#   summarise(n=n())

df.all %>% filter(!is.na(Flag)) %>%
  ggplot(aes(age))+
  geom_histogram(bins=30)+
  facet_grid(.~Flag)+
  theme_minimal()+
  labs(x = "Age", y = "Number of participants")
df.all %>%
  mutate(race = factor(race, levels=c("Boston", "Grandmas", "TC",
   ↪  "Chicago", "NYC"))) %>%
ggplot(aes(x=Wind, y=actual.time/3600, color=sex)) +
  geom_smooth()+
  # scale_y_continuous(breaks = c(3600*c(3,4)),
  #                    labels = c("3HR","4HR"))+
  theme_minimal()+
  labs(x = "Wind speed in Km/hr", y = "Completion Time in hours")+
  facet_grid(.~race)
p.ozone<-df.all %>% ggplot(aes(x=`44201`, y=actual.time/3600,
 ↪  color=age.division)) +
  geom_smooth()+
  # scale_y_continuous(breaks = c(3600*c(3,4)),
  #                    labels = c("3HR","4HR"))+
  theme_minimal()+
  labs(x = "Ozone (parts per million)", y = "Completion Time in hours")+
  facet_grid(.~sex)+
  theme(legend.position="none")

p.pm2.5<-df.all %>% ggplot(aes(x=`88101`, y=actual.time/3600,
 ↪  color=age.division)) +
  geom_smooth()+
  # scale_y_continuous(breaks = c(3600*c(3,4)),
  #                    labels = c("3HR","4HR"))+
  theme_minimal()+
  labs(x = "PM2.5 (micrograms/cubic meter)", y = "Completion Time in
   ↪  hours",
       color = "Age Group")+
  theme(
    legend.text = element_text(size = 5),
    legend.title = element_text(size = 6)
    )+
  facet_grid(.~sex)

grid.arrange(p.ozone, p.pm2.5)
#largest impact analysis: correlation
df.cor.subset<-df.all %>%
  dplyr::select(actual.time, WBGT, Wind, `44201`, `88101`)

colnames(df.cor.subset) <- c("Completion Time", "WBGT", "Wind", "Ozone",
 ↪  "PM2.5")

corrplot.mixed(cor(df.cor.subset, use = "complete.obs"))
df.all <- df.all %>%
  rename(Ozone = `44201`, PM2.5 = `88101`)

df.all$sex<-as.factor(df.all$sex)

#linear model
lm.full <- lm(actual.time ~ WBGT + Wind + PM2.5 +
                Ozone + sex + age, data = df.all)
#summary(lm.full)

#Best Subset Selection: "WBGT Wind Ozone sex age"
#best.subset <- ols_step_best_subset(lm.full)
#best.subset$metrics$predictors[which.min(best.subset$metrics$cp)]
```

```
# #forward selection
# ols_step_forward_aic(lm.full, details=FALSE)
# #backward selection
# ols_step_backward_aic(lm.full, details=FALSE)

lm.best <- lm(actual.time ~ WBGT + Wind + Ozone + sex + age, data =
 ↪  df.all)
s<-summary(lm.best)
kable(coefficients(lm.best), col.names = c("Coefficients"),
      booktabs = TRUE, digits = 3)

adj.r<-s$adj.r.squared
aic<-AIC(lm.best)
```