Project 1: Exploratory Data Analysis

Tianna Chan

Project 1: Exploratory Data Analysis

This project explores different factors that impact marathon performance between gender and across age.

The `project1` data set contains info and results for each participant in the marathon races at Boston, New York City, Chicago, Twin Cities, and Grandmas from 1993 to 2016. It also includes weather parameters, such as dry and wet bulb temperature, relative humidity, black globe temperature and more. This results data already includes calculated variables `WBGT` (Wet Bulb Globe Temperature, calculated by the three temperature) and `Flag` (Groups for WBGT). The project also utilized the `course_record` data that included the record for each race, at each year.

To further assess the environmental impact, AQI data was obtained using provided code in class, which grabbed data from an API using the R package RAQSAPI. The resulted data set includes the average values for ozone level in parts per million, and the PM2.5 in Micrograms/cubic meter (LC).

Data Preprocessing

Missing Data

Project 1 Data: To understand the data set, I first generated a missing data plot by `vis_miss()`. The plot shows that there was a consistent 4% of missing data for the flag and weather measurement columns. This indicates that weather was not measured at these races. After some further exploration, table below shows that all data were missing at races 1 (Chicago), 2 (NYC), 3 (Twin Cities) at 2011, and race 4 (Grandma's) at 2012.

The flag for these were missing because they should have been calculated by the wet bulb globe temperature (WBGT), and the WBGT is also calculated by the other temperature variables, which were not measured and is dependent on these particular races in particular year. Therefore, the probability of the flag and WBGT being missing depends on both observed

and unobserved variables, this should then be a case of **Missing Not at Random (MNAR).** For the other weather variables, they were missing because they are not measured in that particular year/race, but we do not have knowledge on the exact reason. Therefore, the probability of them being missing depends only on observed variables (year/race), they would then be a case of **Missing at Random (MAR).**

Race	Year	n_miss	n_tot
1	2011	126	126
2	2011	131	131
3	2011	118	118
4	2012	116	116

AQI Values Data: Based on the same missing graph plotted, the api column has 16% missing values. Upon further analysis, it was found that all data with sample duration = 1 hour has the api column missing. This is then **Missing At Random (MAR).**

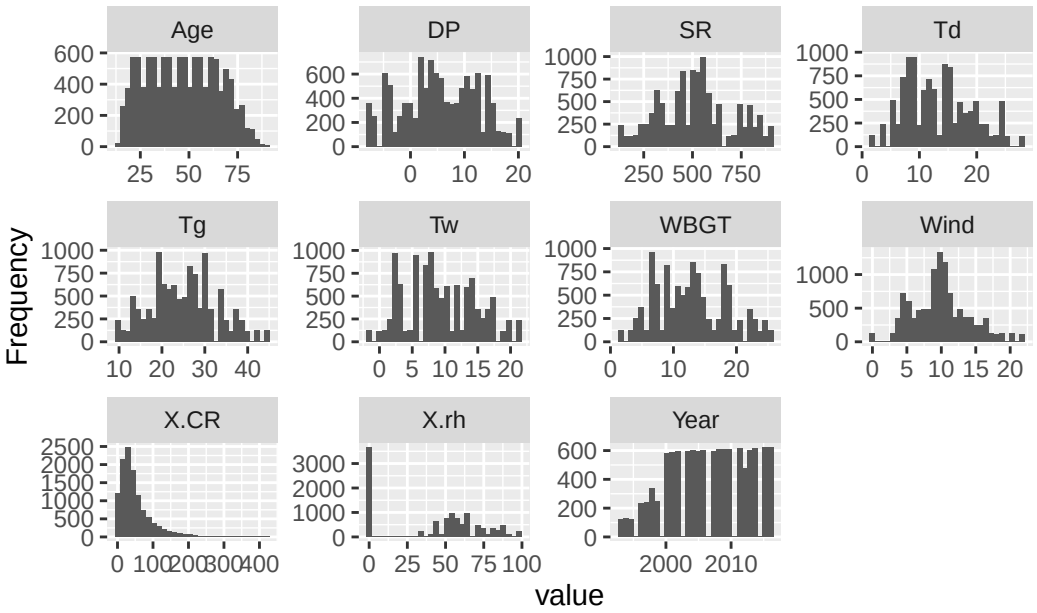
sample_duration	n_miss	n_tot
1 HOUR	1674	1674
24 HOUR	0	4034
24-HR BLK AVG	0	1335
8-HR RUN AVG BEGIN HOUR	0	3408

Course Record Data: No missingness was found in the data.

Data Quality

Project 1 Data: Using `str()`, the variable type for each columns in the data set is shown. The race column was numeric, with values 0, 1, 2, 3, 4. Recoding them into B, C, NY, TC, D would help ease the process of understanding the data and merging with the course record data later. It is further factorized. Similar process for sex, recoding to "M" and "F". Additonally, the Sex and Flag were originally numeric and character variables, and they are also factorized to make sure future analysis treat them the way we want. For the others, they are in the correct numeric value type.

Histograms are plotted for all the numeric values in the data set. Only the %CR (percent off course record) was heavily skewed.



AQI Values Data: Upon glimpsing the data, it was observed that the entries had different coding for marathon races, and it includes the full date for each race. Therefore, recoding was done to ensure later efficiency in merging the data. The summary table below uncover the complexity of this data set, showing the breakdown of information gathered on the data.

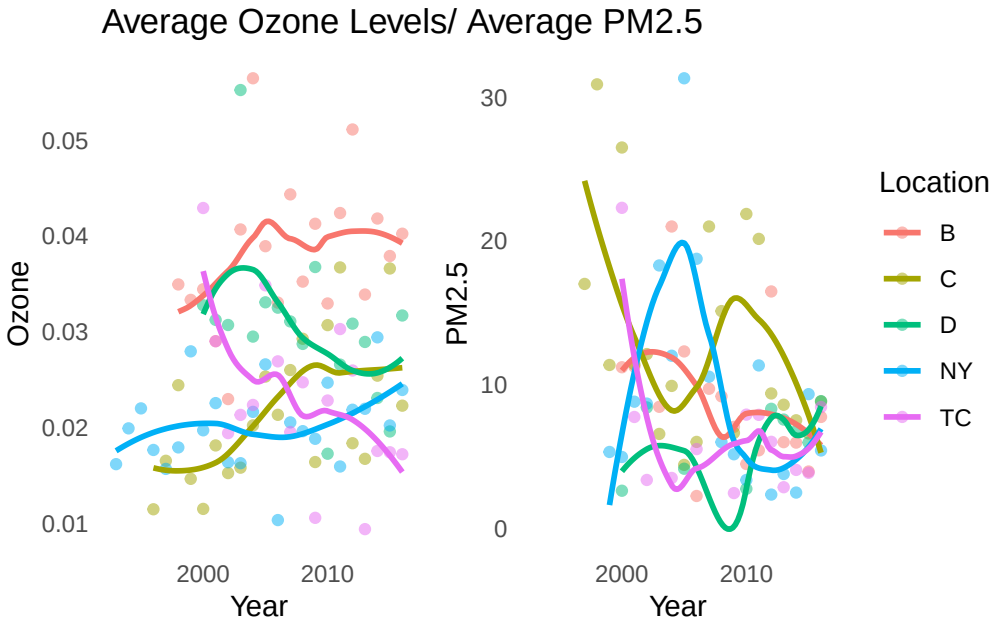
parameter_code	units_of_measure	sample_duration	n
44201	Parts per million	1 HOUR	1136
44201	Parts per million	8-HR RUN AVG BEGIN HOUR	3408
88101	Micrograms/cubic meter (LC)	1 HOUR	124
88101	Micrograms/cubic meter (LC)	24 HOUR	3964
88101	Micrograms/cubic meter (LC)	24-HR BLK AVG	930
88502	Micrograms/cubic meter (LC)	1 HOUR	414
88502	Micrograms/cubic meter (LC)	24 HOUR	70
88502	Micrograms/cubic meter (LC)	24-HR BLK AVG	405

According to the US Environmental Protection Agency, 88101 and 88502 are both used to report daily Air Quality Index values. The difference is that 88101 include both manually operated Federal Reference Methods (FRMs) and automated Federal Equivalent Methods (FEMs), but 88502 are "FRM-like" @EPA. Based on this information, I decide to only take the 88101 data for simplicity. PM2.5 below 12 g/m^3 is considered healthy with little to no risk from exposure, while anything above 35 g/m^3 is unhealthy @IndoorAirHygieneInstitute.

For ozone, it is measured under the parameter 44201. The quality standard is 0.08 ppm @EPA, anything above could be unhealthy. To clean the data, I created a summary by taking the

average of the arithmetic mean.

The graph below on the left illustrates the change in ozone levels on race day, over the years, segmented by marathon locations. Each facet represents a marathon location, allowing for a comparison of trends within those areas. The points represents the average ozone measurements, and we see different fluctuations between different locations. The curve was plotted using `geom_smooth` and shows the trend. In particular, the average ozone levels at New York City was the most stable throughout years; At Twin City we see the greatest decrease; The other locations have a slight fluctuations only. Using the same method, the graph on the right shows the trend on PM2.5. There are big fluctuations for all locations.



Merging Data

The course record data is merged using a `left_join()` so we can use the record to calculate a new variable `Time` for each participant. Since course record has a `Gender` column instead of `Sex`, I have renamed that so I can left join project 1 data to course record data by `Race`, `Year`, and `Sex`. To calculate the time, I used Time (hours) $= \frac{\text{CR} \times \left(1 + \frac{\%CR}{100}\right)}{3600}$ so to adjust for the percent off course record and convert the final unit to hours. Afterwards, the AQI data is joined using `left_join()` as well.

Below shows a correlation plot for all the numeric columns in the final data. There are some high correlations between the weather measures. It is also obvious the course time is highly correlated with age. We will dive deeper and quantify some relations between the variables in the later sections.

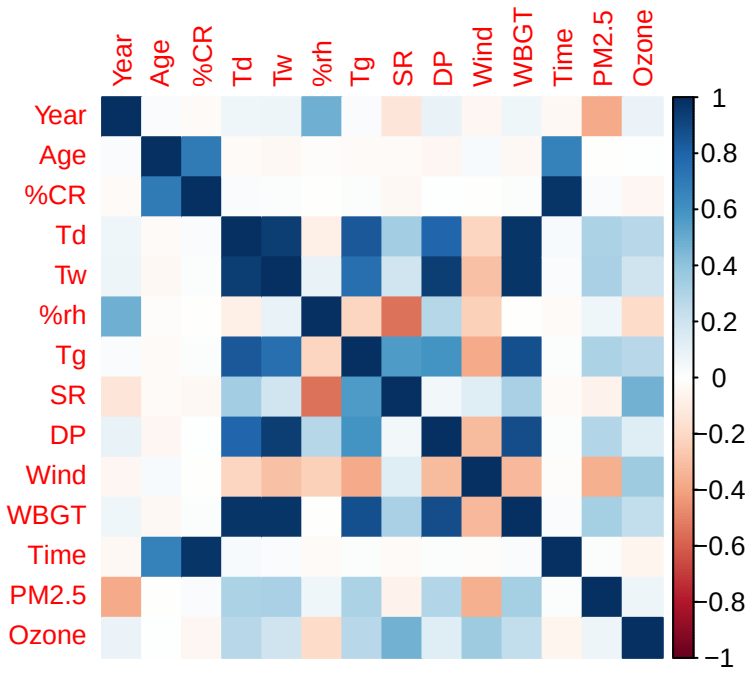
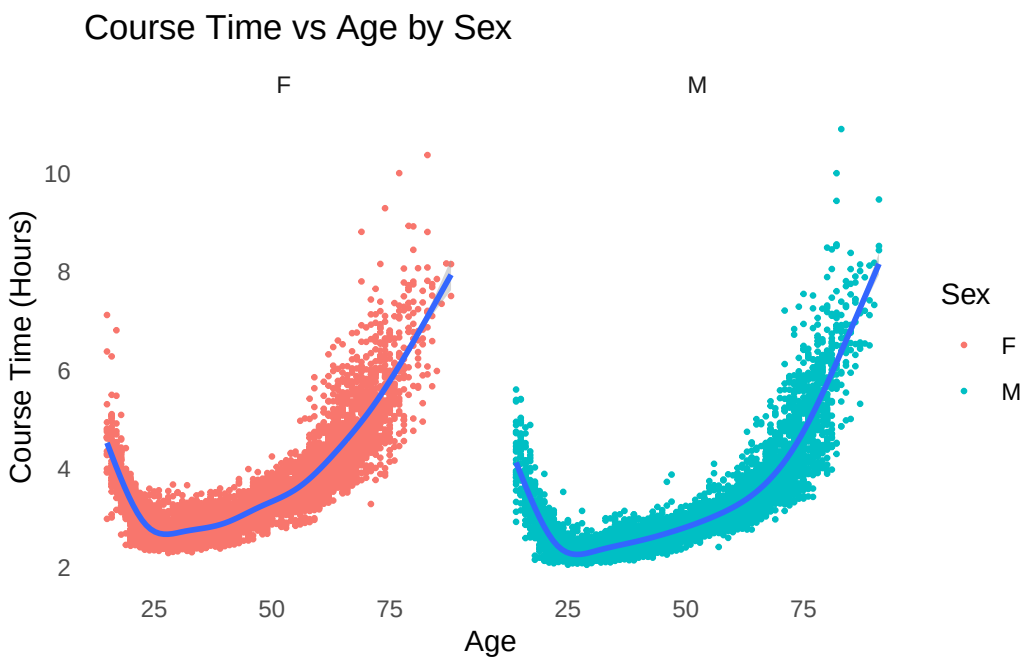


Figure 1: Correlation plot

AIM 1: Examine effects of increasing age on marathon performance in men and women

Plot

The image below shows a scatter plot of time and age for men and women. The plot is divided into two panels, by sex, and the blue line represent a fitted curve performed directly by `geom_smooth()`. There seem to be a near-identical trend for men and women- both sexes gets an increasingly faster time each year they are closer to around age 25. However, as they reach the fastest time at around age 25, they gradually run slower as age increases further. We can observe a spike happening around age 60 for both sexes, but men has a steeper slope. In other words, the slower in time was more significant in the male group. On the other hand, we can see the curve for women is slightly higher than that of the men. This means that women are running slightly slower in general.



This table shows the summary statistics for age and course time for men and women. It is obvious that men (N = 6,112) had a larger sample size than women (N = 5,452). In average, the men population is also older than the female population. We again see that men have a shorter course time on average.

Characteristic	**F**, N = 5,452	**M**, N = 6,112
Age	45 (30, 59)	48 (32, 64)
Time	3.55 (2.87, 3.91)	3.17 (2.49, 3.46)

Regression

A regression model was fit to quantify the course time between sex and their ages. Based on the results below (rounded to 3 decimal places), both Age and Sex are significantly influential on time (p-value < 0.05), with age positively associated while sex negatively associated.

$$E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM}$$

Table 5: Linear Regression Model Summary

Variable	Estimate	Std. Error	t value	Significance
(Intercept)	1.7927	0.0194	92.3411	0

Variable	Estimate	Std. Error	t value	Significance
Age	0.0390	0.0004	104.5268	0
SexM	-0.4909	0.0134	-36.5226	0

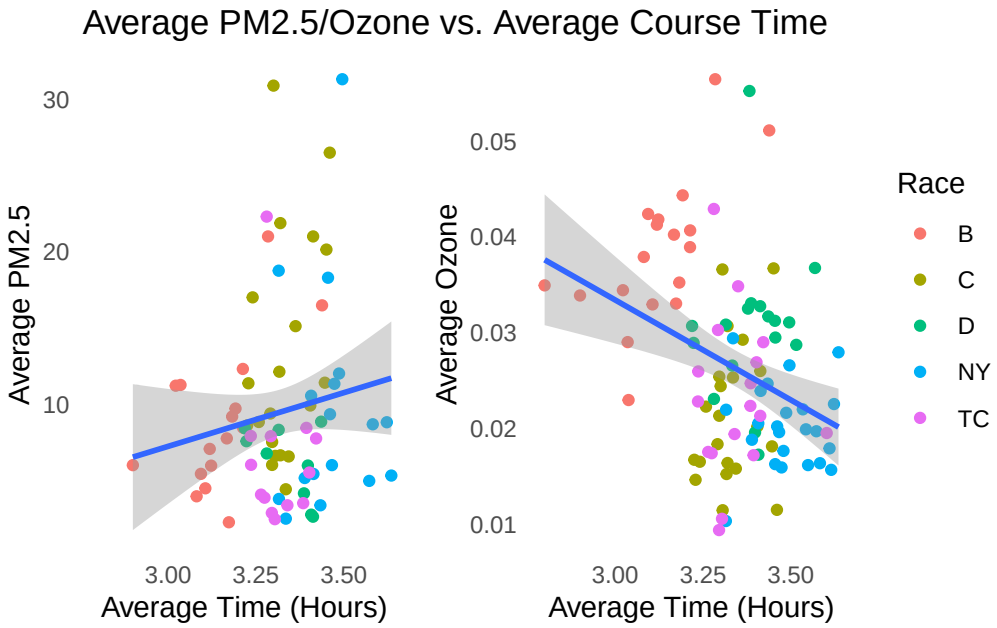
Therefore, for each unit increase in Age, the expected Time increases by $\beta_1 = 0.039$ hours, holding all other variables constant. For **Male** (SexM = 1), the expected Time decreases by $\beta_2 = -0.491$ hours compared to females, holding all other variables constant.

AIM 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

Plots

Most races had pm2.5 on the lower end, under 13. Just by looking at the graph, it does not immediately appear that there is a big difference in performance comparing different pm2.5 levels. The slope drawn by `geom_smooth()` has a tiny increase as average time increases. This suggests that as pm2.5 level goes up, the average course time goes up. In other words, they are running slower on average.

Repeating the same process for Ozone measurements. The slope plotted with `geom_smooth` shows a decreasing trend. This is strange because it suggests that lower ozone may be associated with a longer course time. In other words, runners tends to finish the race faster when the ozone level is higher. However, it is also noticeable from both graphs that the Boston Marathon has a lower average course time among all other marathon races.



Regression

To quantify the impact of environmental conditions on marathon performance, I will first focus on a regression model with only PM2.5 and Ozone: $E[\text{Time}] = \beta_0 + \beta_1 \text{PM2.5} + \beta_2 \text{Ozone}$. Since both PM2.5 and Ozone are significant, both of them were able to explain course time.

$\beta_1 = 0.004$ means that for every one unit increase in PM2.5, the course time increase by 0.004 hours (14.4 seconds), holding others constant.

$\beta_2 = -5.630$ means that for every one unit increase in Ozone, the course time decrease by -5.630 hours. Thus, by conversion, for every 0.01 unit increase in Ozone, the course time decrease by -0.0563 hours (3.378 minutes).

Table 6: Linear Regression Model Summary

Variable	Estimate	Std. Error	t value	Significance
(Intercept)	3.433	0.033	104.121	0.000
PM2.5	0.004	0.002	2.436	0.015
Ozone	-5.630	1.090	-5.167	0.000

To find out whether the impact differs across age and gender, we start with adding the main effects of age and gender, along with the two-way interactions of them with each environmental measure. Then, we perform a backward selection and the results are as follows.

The full model:

$$E[\text{Time}] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{SexM} + \beta_3 \text{PM2.5} + \beta_4 \text{Ozone} + \beta_5(\text{Age} \times \text{PM2.5}) + \beta_6(\text{Age} \times \text{Ozone}) + \beta_7(\text{SexM} \times \text{PM2.5}) + \beta_8(\text{SexM} \times \text{Ozone})$$

Final model by backward selection:

$$E[\text{Time}] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{SexM} + \beta_3 \text{PM2.5} + \beta_4 \text{Ozone} + \beta_5(\text{Age} \times \text{PM2.5}) + \beta_6(\text{Age} \times \text{Ozone}) + \beta_7(\text{SexM} \times \text{PM2.5})$$

Since the interaction term for Sex * Ozone was removed, this suggests that the relationship between Ozone and Time does not differ significantly between males and females in the data set.

Table 7: Linear Regression Model Summary

Variable	Estimate	Std. Error	t value	Significance
(Intercept)	1.7536	0.0651	26.9345	0.0000
Age	0.0409	0.0013	31.8361	0.0000
SexM	-0.4372	0.0264	-16.5871	0.0000
PM2.5	-0.0105	0.0034	-3.0985	0.0020
Ozone	4.1944	2.1464	1.9541	0.0507
Age:PM2.5	0.0004	0.0001	5.5247	0.0000
Age:Ozone	-0.2100	0.0428	-4.9104	0.0000
SexM:PM2.5	-0.0041	0.0023	-1.7602	0.0784

AIM 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

Flag and Course Time

I started with the full model, of all variables in the dataset. Then, I used backward selection to find the smallest best model. This method drops one variable each time and test for the significance. Table below shows the coefficients of the final model by backward selection.

$$E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM} + \beta_3 \cdot \text{FlagRed} + \beta_4 \cdot \text{FlagWhite} + \beta_5 \cdot \text{FlagYellow} + \beta_6 \cdot \text{Td} + \beta_7 \cdot \text{Tw} + \beta_8 \cdot \%\text{rh} + \beta_9 \cdot \text{SR} + \beta_{10} \cdot \text{DP}$$

Table 8: Linear Regression Model Summary: Backward Selection

Variable	Estimate	Std. Error	t value	Significance
(Intercept)	1.726	0.065	26.651	0.000
Age	0.039	0.000	102.896	0.000
SexM	-0.492	0.014	-36.187	0.000
FlagRed	0.127	0.046	2.747	0.006
FlagWhite	-0.028	0.025	-1.124	0.261
FlagYellow	0.087	0.027	3.168	0.002
Td	-0.016	0.009	-1.780	0.075
Tw	0.059	0.020	2.939	0.003
`%rh`	-0.001	0.000	-1.918	0.055
SR	0.000	0.000	-4.274	0.000
DP	-0.028	0.009	-3.105	0.002

Therefore,

For **Female** (SexM = 0), each unit increase in time (hour)

$$E[Time] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM} + \beta_3 \cdot \text{Flag} + \beta_4 \cdot \text{Td} + \beta_5 \cdot \text{Tw} + \beta_6 \cdot \%\text{rh} + \beta_7 \cdot \text{Tg} + \beta_8 \cdot \text{SR} + \beta_9 \cdot \text{DP} + \beta_{10} \cdot \text{Wind} + \beta_{11} \cdot \text{WBGT}$$

Code Appendix

```
library(readr)
library(tidyverse)
set.seed(123456)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(readr)
library(visdat)
library(naniar)
library(gtsummary)
library(patchwork)
library(DataExplorer)
library(lubridate) #To convert time
# Import data sets
aqi_values <- read_csv("aqi_values.csv")
course_record <- read_csv("course_record.csv")
marathon_dates <- read_csv("marathon_dates.csv")
project1 <- read_csv("project1.csv")

colnames(project1) <- c("Race", "Year", "Sex", "Flag", "Age", "%CR", "Td",
 ↪  "Tw", "%rh", "Tg", "SR", "DP", "Wind", "WBGT")
## Missing data plot
# project1 %>% abbreviate_vars() %>% vis_miss()

## Missing data table
project1 %>% group_by(Race, Year) %>%
  summarize(n_miss = sum(is.na(Flag)),
            n_tot = n()) %>%
  filter(n_miss > 0) %>%
  kable(digits = 2)

## AQI Values Data
# vis_miss(aqi_values)
aqi_values %>% select(c(sample_duration, aqi)) %>%
  group_by(sample_duration) %>%
  summarise(n_miss = sum(is.na(aqi)),
            n_tot = n()) %>%
  kable(digits = 2)
```

```r
## Course Record Data
# vis_miss(course_record)
## Data Quality
# str(project1)

project1 <- project1 %>%
  mutate(Race = case_when(
    Race == 0 ~ "B",
    Race == 1 ~ "C",
    Race == 2 ~ "NY",
    Race == 3 ~ "TC",
    Race == 4 ~ "D",
    ),
    Sex = case_when(
    Sex == 0 ~ "F",
    Sex == 1 ~ "M"
    )
    ) %>%
  mutate(across(c(Race, Sex, Flag),
         as.factor))
plot_histogram(project1, nrow = 5L)
# plot_bar(project1, ncol = 5L)
aqi_values %>%
  group_by(parameter_code, units_of_measure, sample_duration) %>%
  summarise(n=n()) %>%
  kable()

# Chnange marathon names and add year column
aqi_values_adj <- aqi_values %>%
  mutate(
    marathon = case_when(
      marathon == "Boston" ~ "B",
      marathon == "Chicago" ~ "C",
      marathon == "NYC" ~ "NY",
      marathon == "Twin Cities" ~ "TC",
      marathon == "Grandmas" ~ "D"
      ),
    year = year(date_local)
    )
aqi_values_adj <- aqi_values_adj %>%
  filter(parameter_code %in% c(88101, 44201)) %>%
    ↪ select(-c("cbsa_code","state_code","county_code","site_number","date_local"))
```

```
aqi_summary_by_year <- aqi_values_adj %>%
  group_by(year, marathon, units_of_measure) %>%
  summarise(avg_arithmetic_mean=mean(arithmetic_mean))

aqi_summary_by_year <-
  spread(aqi_summary_by_year, units_of_measure, avg_arithmetic_mean) %>%
  arrange(year)

colnames(aqi_summary_by_year) <- c("Year", "Race", "PM2.5", "Ozone")
a <- ggplot(aqi_summary_by_year, aes(x = Year, y = Ozone, color = Race)) +
  geom_point(alpha=0.5) +
  geom_smooth(se = FALSE) +
  # facet_wrap(~ Race) +
  labs(title = "Average Ozone Levels/ Average PM2.5",
       x = "Year") +
  theme_minimal() +
  scale_color_discrete(name = "Location") +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  guides(color="none")
b <- ggplot(aqi_summary_by_year, aes(x = Year, y = `PM2.5`, color = Race)) +
  geom_point(alpha=0.5) +
  geom_smooth(se = FALSE) +
  # facet_wrap(~ Race) +
  labs(x = "Year") +
  theme_minimal() +
  scale_color_discrete(name = "Location") +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
a+b
course_record <- course_record %>% rename("Sex" = "Gender")

project1_CR <- project1 %>%
  left_join(course_record, by = c("Race", "Year", "Sex"))

project1_CR$Time <- (project1_CR$CR * (1+project1_CR$`%CR`/100))/3600 %>%
  as.numeric()
project1_CR_aqi <- project1_CR %>%
  left_join(aqi_summary_by_year, by = c("Race", "Year"))
library(corrplot)

project1_CR_aqi$Time <- project1_CR_aqi$Time %>% as.numeric()
```

```r
data <- select_if(project1_CR_aqi,is.numeric) %>% na.omit() %>%
 ↪  abbreviate_vars()
M = cor(data)
corrplot(M, method = 'color', tl.cex = 0.8)
ggplot(project1_CR_aqi, aes(x = Age, y = Time)) +
  geom_point(aes(color=Sex), size = 0.5) +
  geom_smooth() +
  facet_wrap(~ Sex) +
  labs(title = "Course Time vs Age by Sex",
       x = "Age",
       y = "Course Time (Hours)") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
project1_CR_aqi %>% dplyr::select(Age, Sex, Time)%>%
  tbl_summary(by = Sex,
              type = list(where(is.numeric) ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({p25}, {p75})"))
               ↪  %>%
  bold_labels() #%>%
  #add_p()
## Time effects on Sex and Age
model <- lm(as.numeric(Time) ~ Age + Sex, data = project1_CR_aqi)
summary <- summary(model)

# Presenting results in table format
coefficients <- as.data.frame(summary$coefficients)

kable(coefficients, digits = 4,
      caption = "Linear Regression Model Summary",
      col.names = c("Variable", "Estimate", "Std. Error", "t value",
       ↪  "Significance"))
new_data <- project1_CR_aqi %>% group_by(Race, Year, PM2.5, Ozone) %>%
 ↪  summarize(Avg_Time = mean(as.numeric(Time)))

a <- ggplot(new_data, aes(x = Avg_Time, y = PM2.5)) +
  geom_point(aes(color = Race)) +
  geom_smooth(method = "lm") +
  labs(title = "Average PM2.5/Ozone vs. Average Course Time",
       x = "Average Time (Hours)",
       y = "Average PM2.5") +
  theme_minimal() +
```

```
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  guides(color="none")
b <- ggplot(new_data, aes(x = Avg_Time, y = Ozone)) +
  geom_point(aes(color = Race)) +
  geom_smooth(method = "lm") +
  labs(x = "Average Time (Hours)",
       y = "Average Ozone") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

a + b
model <- lm(Time ~ PM2.5 + Ozone, data = project1_CR_aqi)
summary <- summary(model)

coefficients <- as.data.frame(summary$coefficients)

kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary",
      col.names = c("Variable", "Estimate", "Std. Error", "t value",
       ↪  "Significance"))
full_model <- lm(Time ~ Age + Sex + PM2.5 + Ozone
                 + Age * PM2.5 + Age * Ozone
                 + Sex * PM2.5 + Sex * Ozone,
                 data = project1_CR_aqi)

# summary(full_model)

backward_model <- step(full_model, direction = "backward")
summary <- summary(backward_model)

coefficients <- as.data.frame(summary$coefficients)
kable(coefficients, digits = 4,
      caption = "Linear Regression Model Summary",
      col.names = c("Variable", "Estimate", "Std. Error", "t value",
       ↪  "Significance"))
# ggplot(project1_CR_aqi$Time, col = project1_CR_aqi$Flag)
## Backward selection models

# Full model
model <- lm(as.numeric(Time) ~ Age + Sex + Flag + Td + Tw + `%rh` + Tg + SR +
 ↪  DP + Wind + WBGT, data = project1_CR_aqi)
```

```r
# summary(model)

# Backward selection model
final_model <- step(model, direction = "backward")
summary <- summary(final_model)

# Presenting results in table format
coefficients <- as.data.frame(summary$coefficients)

kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary: Backward Selection",
      col.names = c("Variable", "Estimate", "Std. Error", "t value",
       ↪  "Significance"))
```