Project 1: Exploratory Data Analysis

Due: October 6th at 11:59pm

Daniel Posmik (daniel_posmik@brown.edu)

Overview

This project is a result of a collaboration with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. Endurance exercise performance is degraded with increasing environmental temperature, and the decline in performance associated with warmer temperatures is magnified with longer-distance events such as the marathon footrace (42.2k). In addition, older adults experience thermoregulatory challenges that impair their ability to dissipate heat, which can further exacerbate the impact of warmer temperatures. Finally, there are well-documented sex differences in endurance performance and in physiological processes related to thermoregulation. The purpose of Dr. Ely's research is to examine the impact of environmental conditions including temperature, humidity, solar radiation, and wind on marathon performance in men and women throughout the lifespan. This data set includes top single-age performances from five major marathons across 15-20 years from age 14-85 in men and women, with detailed environmental conditions for each marathon.

During this project, the focus will be on three aims:

- Aim 1: Examine effects of increasing age on marathon performance in men and women
- Aim 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.
- Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

Data Pre-Processing

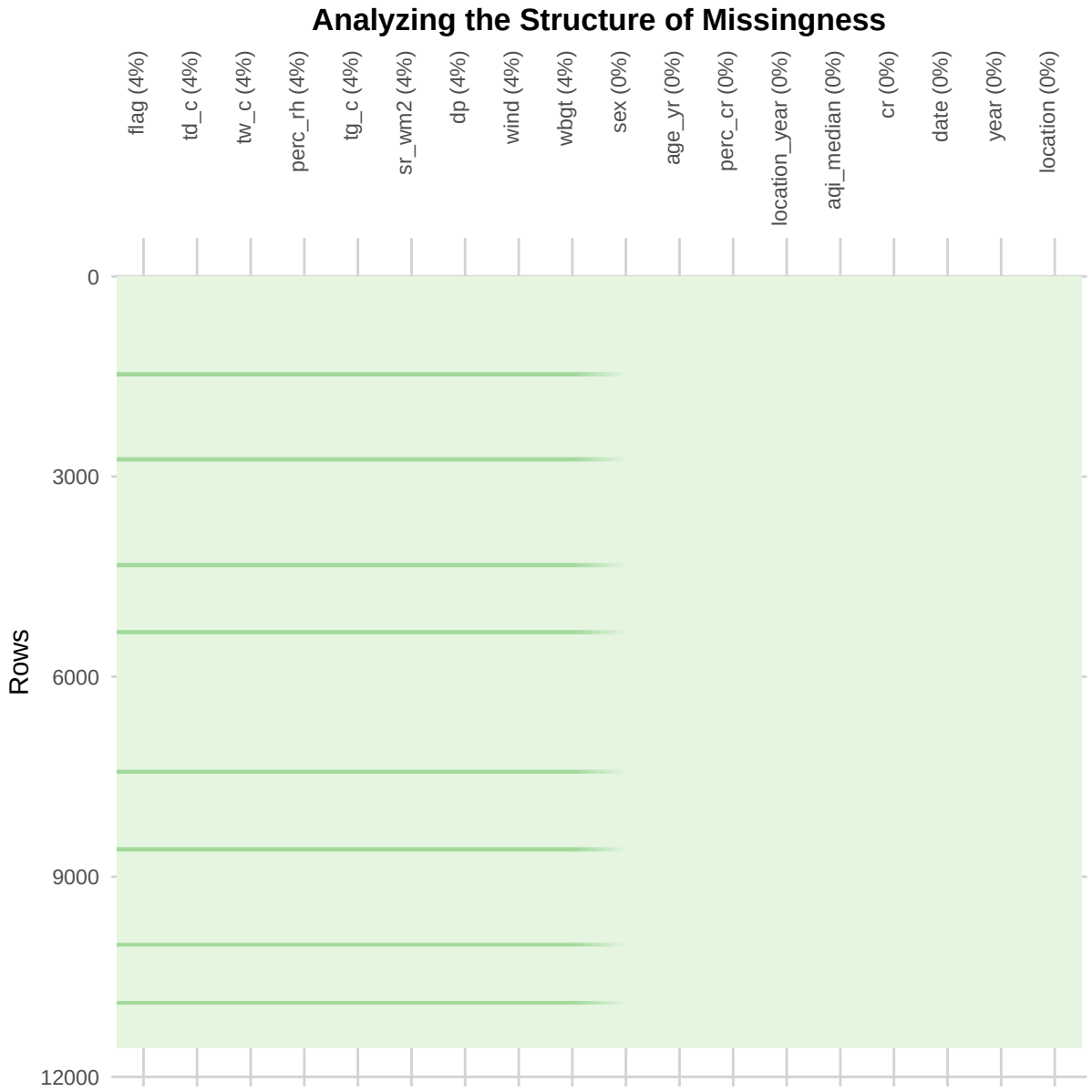
First, let us combine the data from the different datasets into a single dataset. Since each city has one marathon per year, we can create a unique identifier for each marathon by combining the city and the year. We will then use the "project1.csv' dataset, containing the individual-level data, and merge the other datasets onto it.

Particular care must be taken when merging in the AQI data (`aqi_values`). The aqi values themselves are reported in two different units: `ppm` (n=4,544) and `ug/m3` (n=5,907) so we need to use the arithmetic mean as a composite measure for our air quality measure. Since there are multiple point measurements for each location and year, we will take the median aqi value for each location and year.

Table 1: Variable Summary for main_dta

Variable Name	Variable Type	Description
sex	numeric	Sex of the runner (0 = Female, 1 = Female)
flag	factor	Flag: Based on WGBT and risk of heat illness
age_yr	numeric	Age in years
perc_cr	numeric	Percentage off course record
td_c	numeric	Dry bulb temperature in celsius
tw_c	numeric	Wet bulb temperature in celsius
perc_rh	numeric	Percent relative humidity
tg_c	numeric	Black globe temperature in celsius
sr_wm2	numeric	Solar radiation in watts per meter squared
dp	numeric	Dew point in celsius
wind	numeric	Wind speed in km/h
wbgt	numeric	Wet bulb globe temperature
location_year	character	Race location and year identifier
aqi_median	numeric	Median AQI value (Derived from arithmetic mean measure)
cr	numeric	Course record time in seconds
date	Date	Race date YYYY-MM-DD
year	numeric	Race year
location	character	Race location

Additionally, let us conduct some preliminary data integrity checks and assess missingness.



There is a striking pattern of missingness in the data. The data show a row-specific pattern of missingness that re-occurs regularly. Interestingly, if a variable has missing values, it has exactly 491 missing values ($\sim$ 4%). This suggests that the missingness is not random, but rather systematic.

Table 2: Count of Missingness by Year and Location

year	NY	C	B	D	TC
1993	0	0	0	0	0
1994	0	0	0	0	0
1995	0	0	0	0	0
1996	0	0	0	0	0
1997	0	0	0	0	0
1998	0	0	0	0	0
1999	0	0	0	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	1179	1134	0	0	1062
2012	0	0	0	1044	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

A glance at missingness by year and location reveals that missingness mostly stems from the year 2011 with another chunk of missingness the 2012 "Grandmas" Duluth marathon. The Boston marathon is the only race unaffected by this missingness.

Aim 1: Examine effects of increasing age on marathon performance

To examine the effects of increasing age on marathon performance, we construct our outcome variable `completion_time` = `perc_cr` $\cdot$ `cr` (i.e. "percentage off course record" times "course record") and use age in years (`age_yr`) as our predictor. Let us first plot this relationship, paying attention to gender.

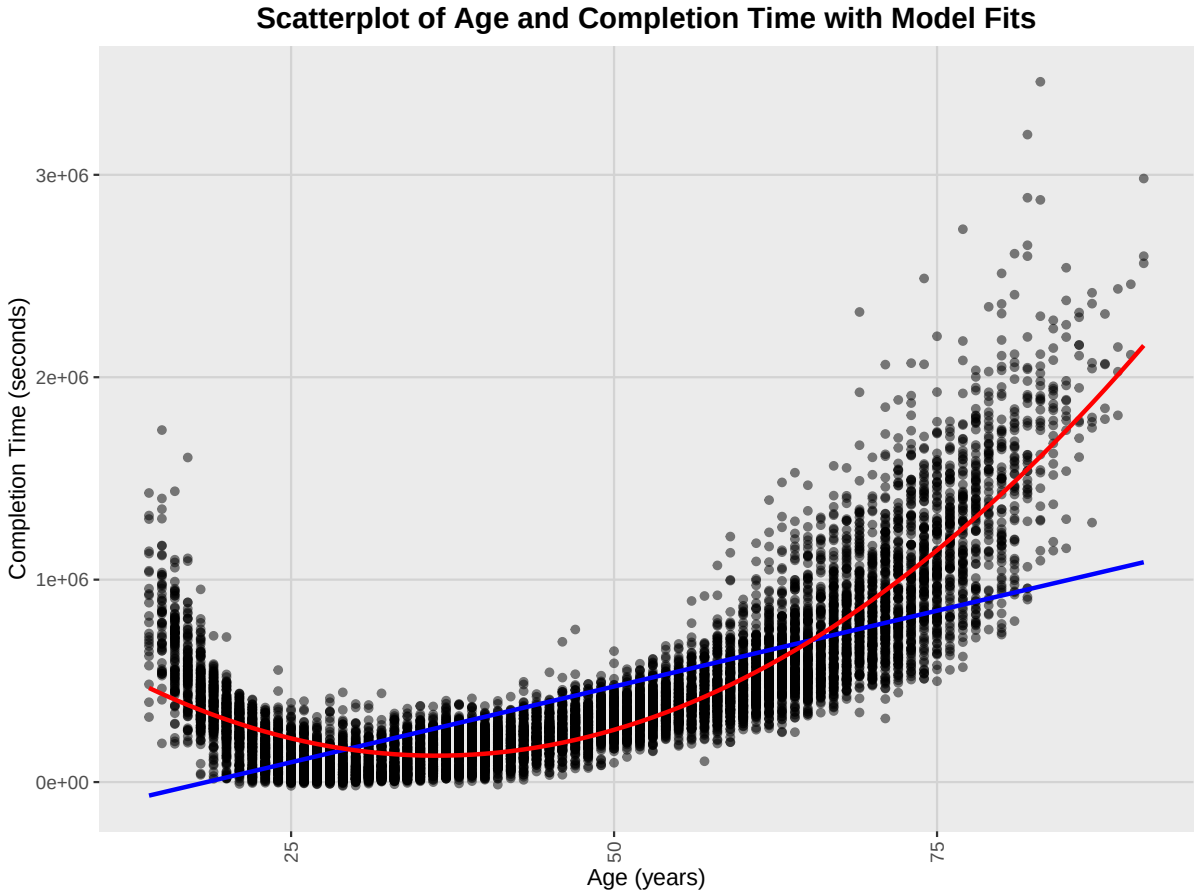
The scatterplot shows a non-linear relationship between age and completion time. Until roughly age 25, completion time decreases. From age 25 to age 35, completion time remains low and steady. After 35, completion time increases. This is not a surprising trend, seeing that the human body peaks in performance around the mid-20s. Furthermore, the data show slight heteroskedasticity, with the variance of completion time increasing with age. The data show us that the older the runner, the more diverse the completion times.

Moreover, differences between genders become visible when age increases. It seems that female completion times rise sooner and more sharply than male completion times. This may be a result of more male runners participating in marathons at older ages. All in all, the differences between genders are not as pronounced as the differences between age groups.

In order to model the relationship between age and marathon performance, we will employ an ordinary least squares regression model, contrasting how a squared term of age (`age_yr_sq`) affects the model fit. Since the focus of this project is on exploratory analysis, we will not delve into the specifics of model and variable (i.e. controls) selection. That being said, we can always find methods for modeling these relationships that are more complex. More advanced choices may include spline regression, kernel-based regression, or other unsupervised learning

methods. Since the data are relatively straightforward, we will stick to ordinary least squares regression.

Now, we fit two models: one with a linear term for age and one with a quadratic term for age and visualize the results for convenience using the `modelsummary` package.



We can see that the linear regression model with the quadratic term fits the data better than the simple linear model. The quadratic model captures the non-linear relationship between age and completion time. The quadratic model shows that completion time decreases until roughly age 25, remains low and steady until age 35, and then increases. The linear model, on the other hand, does not capture the non-linear relationship between age and completion time. That being said, the model with the quadratic term is by no means perfect. We can see that the model does not capture the variance in completion time well as age increases.

All in all, if we trust the model with the quadratic term, we can say that–all else equal–for x additional years in age, completion time changes by $-49003 \cdot x + 676 \cdot x^2$ seconds.

	Linear Model (Simple)	Linear Model (Quadratic)
(Intercept)	−276 423.027	1 018 509.066
	(7103.838)	(11 049.031)
age_yr	14 972.762	−49 003.303
	(142.558)	(505.645)
age_yr_sq		675.940
		(5.254)
Num.Obs.	11 564	11 564
R2	0.488	0.790
R2 Adj.	0.488	0.789
AIC	322 534.4	312 261.8
BIC	322 556.4	312 291.2
Log.Lik.	−161 264.183	−156 126.887
F	11 031.191	21 684.512
RMSE	275 522.59	176 694.21

Aim 2: Explore the impact of environmental conditions on marathon performance

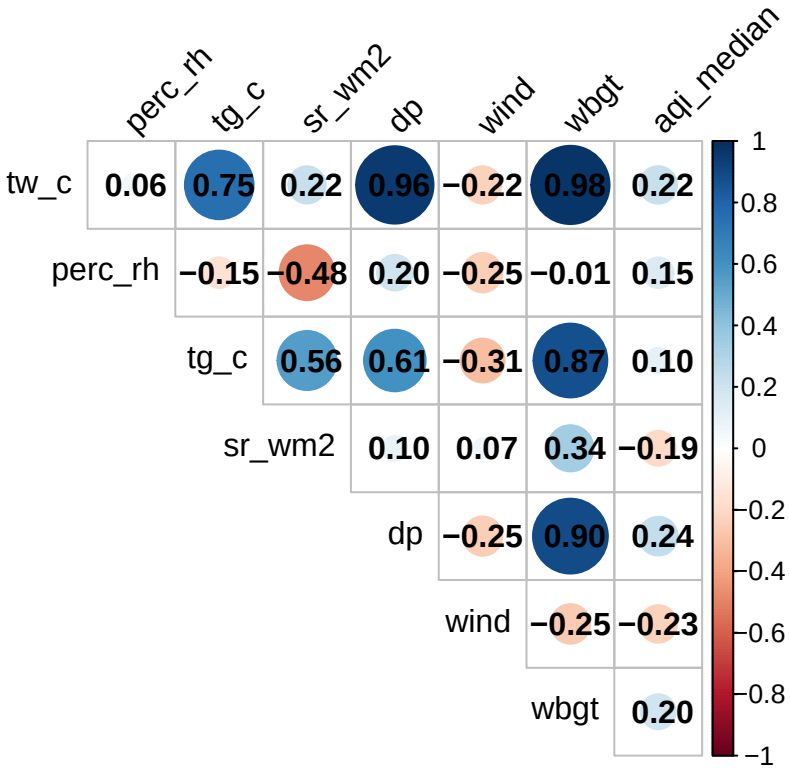
Next, we are interested in the impact of environmental conditions on marathon performance. To start, let us examine some summary statistics of the weather variables by location.

Table 3: Summary of Weather Variables by Location

location	variable	n	mean	sd	min	max	range
B	tw_c	2088	8	4	2	18	15
B	perc_rh	2088	36	34	0	98	98
B	tg_c	2088	24	8	10	42	33
B	sr_wm2	2088	650	187	147	853	706
B	dp	2088	3	4	-4	14	18
B	wind	2088	12	4	5	22	17
B	wbgt	2088	11	5	7	23	17
B	aqi_median	2088	3	3	0	10	10
C	tw_c	2427	9	6	-1	21	23
C	perc_rh	2427	60	10	43	85	42
C	tg_c	2427	25	6	10	36	25
C	sr_wm2	2427	460	95	253	608	356

location	variable	n	mean	sd	min	max	range
C	dp	2427	5	7	-7	20	27
C	wind	2427	8	3	3	16	13
C	wbgt	2427	12	6	1	25	23
C	aqi_median	2553	4	6	0	17	17
D	tw_c	1884	15	2	10	20	10
D	perc_rh	1884	49	34	0	90	89
D	tg_c	1884	32	8	14	44	31
D	sr_wm2	1884	677	191	289	909	620
D	dp	1884	12	3	4	18	14
D	wind	1884	9	3	4	14	10
D	wbgt	1884	19	3	14	25	11
D	aqi_median	2000	2	3	0	8	8
NY	tw_c	2799	8	5	1	17	16
NY	perc_rh	2799	27	30	0	98	98
NY	tg_c	2799	21	6	11	35	23
NY	sr_wm2	2799	401	131	143	573	431
NY	dp	2799	3	7	-7	16	23
NY	wind	2799	11	5	0	20	20
NY	wbgt	2799	11	5	4	19	15
NY	aqi_median	2930	4	7	0	31	31
TC	tw_c	1875	10	5	2	22	20
TC	perc_rh	1875	42	34	0	89	89
TC	tg_c	1875	25	7	13	35	23
TC	sr_wm2	1875	436	139	141	630	489
TC	dp	1875	6	7	-7	20	28
TC	wind	1875	9	3	4	16	12
TC	wbgt	1875	13	5	7	24	18
TC	aqi_median	1993	6	5	0	22	22

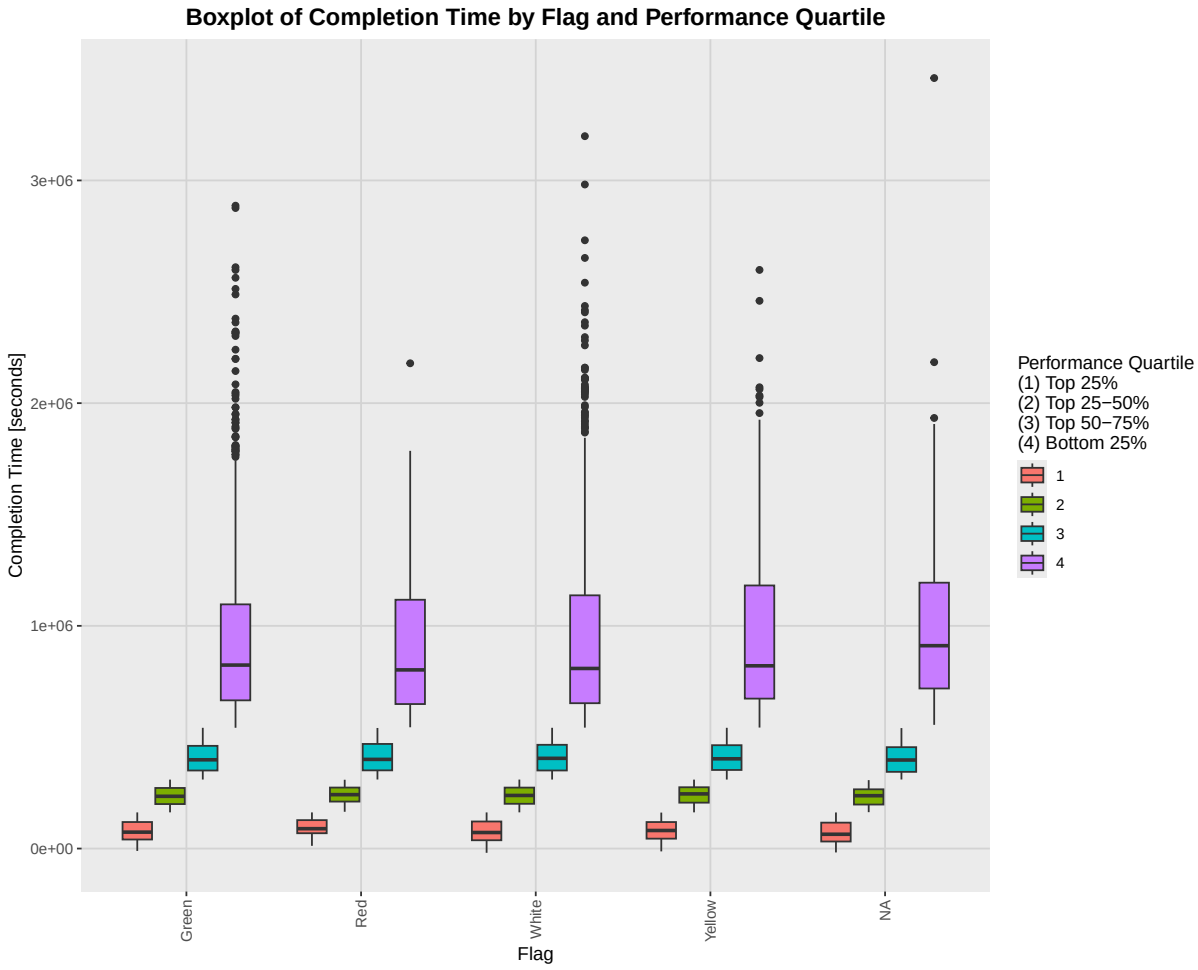
From this table, it becomes apparent that the "Grandmas" Duluth marathon has a slightly higher temperature mean than the other marathons. The "Grandmas" Duluth marathon also has the highest solar radiation and dew point mean. This suggests that the "Grandmas" Duluth marathon is the "warmest" marathon. Now, when analyzing which weather variables have the largest impact on marathon performance, we have to take correlations into account. This is especially important if the analyst attempts to model the relationship between weather variables and marathon performance. High correlations between explanatory variables can lead to multicollinearity, which can cause estimation issues. That is why we will first examine the correlation matrix of the weather variables.



We can immediately see that there are strong correlation within the weather variables. For example, `wbgt` (wet bulb globe temperature) is highly correlated with `tw_c` (temperature in Celsius) and `dp` (dew point). Of course, some of these strong correlations are to be expected, as WGBT is derived from temperature.

Now, let us turn to the relationship between completion time and weather variables. We will start by examining the relationship between completion time and the weather variable `flag`. The `flag` variable is based on the Wet Bulb Globe Temperature (WBGT) and the risk of heat illness. Exploratory analysis that is not reported suggests that the curvature of the best fit curve is similar across the flag categories. A more notable difference is the intercept of the regression line. This suggests that the impact of the flag on completion time is relatively constant across different ages. That being said, it remains interesting to see whether the impact of weather (i.e. `flag`) on completion time differs across performance groups. This hypothesis could be based on the assumption that less strong runners are more susceptible to heat effects.

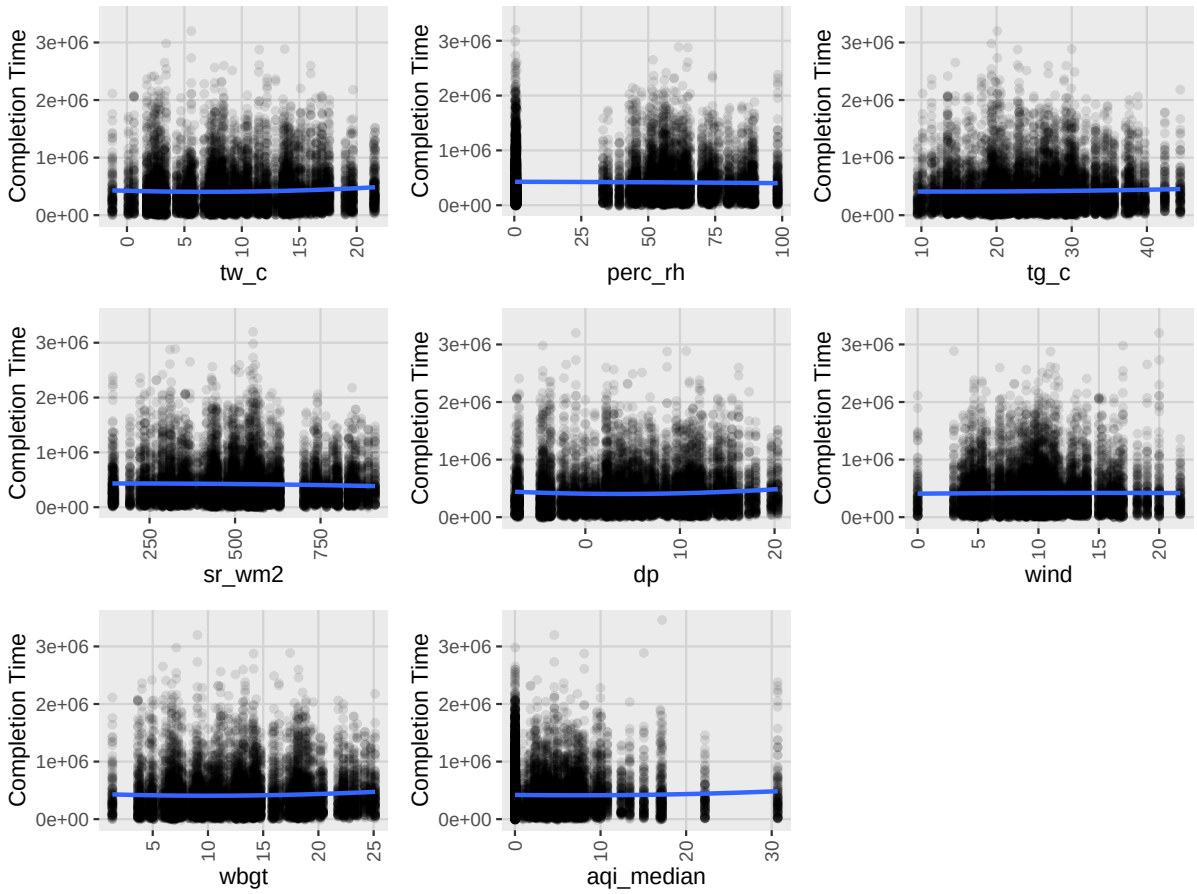
We will now classify runners into performance quartiles and examine the relationship between completion time and flag.



Despite accounting for the performance quartiles, there is little to no noticeable impact of the flag category on completion time. This suggests that less strong runners are in fact not more susceptible to the variations in weather conditions.

Aim 3: Identify the weather parameters that have the largest impact on marathon performance

Lastly, we hope to identify the weather parameters that have the largest impact on marathon performance. To begin, let us visualize the relationship between completion time and the weather variables with a scatterplot matrix and a line of best fit.



The scatterplots reveal no significant relationships. A lot of the scatterplots show a flat line of best fit and bunching of x-values around discrete values like 0, suggesting that the weather variables do not have a strong impact on completion time.

To definitively measure which weather parameters have the largest impact on marathon performance, we will conduct a regression analysis. We will regress completion time on a subset of weather variables and examine the coefficients. The previous correlation analysis will guide our selection of weather variables.

```
Call:
lm(formula = completion_time ~ tw_c + perc_rh + tg_c + sr_wm2 +
    dp + wind + wbgt + aqi_median, data = main_dta)

Residuals:
    Min      1Q  Median      3Q     Max
-491983 -255313 -107175  125651 2761470
```

	(1)
(Intercept)	431 549.562
	(29 804.775)
tw_c	−48 143.706
	(45 667.899)
perc_rh	−423.430
	(153.862)
tg_c	−13 627.546
	(10 437.564)
sr_wm2	−189.408
	(30.187)
dp	−4203.574
	(4864.392)
wind	597.291
	(1095.146)
wbgt	70 898.002
	(50 708.049)
aqi_median	−657.520
	(756.961)
Num.Obs.	11 073
R2	0.007
R2 Adj.	0.007
AIC	316 120.6
BIC	316 193.7
Log.Lik.	−158 050.308
F	10.337
RMSE	382 524.23

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 431549.56   29804.77  14.479  < 2e-16 ***
tw_c        -48143.71   45667.90  -1.054  0.29181
perc_rh       -423.43     153.86  -2.752  0.00593 **
tg_c        -13627.55   10437.56  -1.306  0.19171
sr_wm2         -189.41      30.19  -6.274 3.64e-10 ***
dp            -4203.57    4864.39  -0.864  0.38752
wind            597.29    1095.15   0.545  0.58549
wbgt          70898.00   50708.05   1.398  0.16209
aqi_median     -657.52     756.96  -0.869  0.38507
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382700 on 11064 degrees of freedom
  (491 observations deleted due to missingness)
Multiple R-squared:  0.007419,  Adjusted R-squared:  0.006701
F-statistic: 10.34 on 8 and 11064 DF,  p-value: 1.595e-14
```

Code Appendix

```r
# Set up knit environment
knitr::opts_chunk$set(echo = F)
knitr::opts_chunk$set(error = F)
knitr::opts_chunk$set(warning = F)
knitr::opts_chunk$set(message = F)

# Load necessary packages
library(tidyverse)
library(magrittr)
library(lubridate)
library(GGally)
library(corrplot)
library(kableExtra)
library(gridExtra)
library(knitr)
library(ggplot2)
library(naniar)
library(gtsummary)
```

```r
# Define folders
base_folder <-
  "/Users/posmikdc/Documents/brown/classes/php2550-pda/"

input_folder <-
  paste0(base_folder, "php2550-projects/project1/data/")

output_folder <-
  paste0(base_folder, "php2550-projects/project1/output/")

# Load data
project1_dta <- read_csv(paste0(input_folder, "project1.csv"))
course_record <- read_csv(paste0(input_folder, "course_record.csv"))
marathon_dates <- read_csv(paste0(input_folder, "marathon_dates.csv"))
aqi_values <- read_csv(paste0(input_folder, "aqi_values.csv"))

# Create a year variable for aqi_values
aqi_values %<>%
  mutate(year =
    lubridate::year(as.Date(date_local))
  )

# Create a crosswalk for the marathon location
xwalk_location <- as.data.frame(
  list(
    project1_id = c("0", "1", "2", "3", "4"),
    marathon_dates_id = c("Boston", "Chicago", "NYC", "Twin Cities", "Grandmas"),
    course_record_id = c("B", "C", "NY", "TC", "D"),
    aqi_id = c("Boston", "Chicago", "NYC", "Twin Cities", "Grandmas")
  )
)

# Create location-year identifiers
project1_dta %<>%
  rename(race = "Race (0=Boston, 1=Chicago, 2=NYC, 3=TC, 4=D)") %>%
  mutate(race = as.character(race)) %>%  # Convert race to character
  left_join(xwalk_location %>% select(project1_id, course_record_id),
            by = c("race" = "project1_id")) %>%
  rename(location = course_record_id) %>%
  mutate(location_year = paste0(location, "_", Year))

course_record %<>%
```

```
  rename(location = Race) %>%
  mutate(location_year = paste0(location, "_", Year))

marathon_dates %<>%
  left_join(xwalk_location %>% select(marathon_dates_id, course_record_id),
            by = c("marathon" = "marathon_dates_id")) %>%
  rename(location = course_record_id) %>%
  mutate(location_year = paste0(location, "_", year))

aqi_values %<>%
  left_join(xwalk_location %>% select(aqi_id, course_record_id),
            by = c("marathon" = "aqi_id")) %>%
  rename(location = course_record_id) %>%
  mutate(location_year = paste0(location, "_", year))

# Clean aqi_values
aqi_trim <- aqi_values %>%
  select(location_year, arithmetic_mean) %>%
  group_by(location_year) %>%
  summarize(
    aqi_median = median(arithmetic_mean, na.rm = TRUE)
  )

# Merge data datasets with project1_dta
main_dta <- project1_dta %>%
  left_join(aqi_trim, by = "location_year") %>%
  left_join(course_record, by = "location_year") %>%
  left_join(marathon_dates, by = "location_year") %>%
  select(- c(race, Year.x, location.x, location.y, Year.y, marathon)) %>%
  mutate(
    Flag = as.factor(Flag),
    CR = lubridate::time_length(CR, "seconds")
    ) %>%
  rename_with(~ gsub("[,()/]", "", .)) %>%
  rename_with(tolower, everything()) %>%
  rename_with(str_replace_all, pattern = " ", replacement = "_") %>%
  rename_with(str_replace_all, pattern = "%", replacement = "perc_") %>%
  rename(sex = "sex_0=f_1=m")

# Write intermediate dataset
write_csv(main_dta, paste0(output_folder, "main_dta.csv"))
```

```r
# Summary table
table_summary <- tibble(
  "Variable Name" = colnames(main_dta),
  "Variable Type" = sapply(main_dta, class),
  "Description" = c(
    "Sex of the runner (0 = Female, 1 = Female)", #1
    "Flag: Based on WGBT and risk of heat illness", #2
    "Age in years", #3
    "Percentage off course record", #4
    "Dry bulb temperature in celsius", #5
    "Wet bulb temperature in celsius", #6
    "Percent relative humidity", #7
    "Black globe temperature in celsius", #8
    "Solar radiation in watts per meter squared", #9
    "Dew point in celsius", #10
    "Wind speed in km/h", #11
    "Wet bulb globe temperature", #12
    "Race location and year identifier", #13
    "Median AQI value (Derived from arithmetic mean measure)", #14
    "Course record time in seconds", #15
    "Race date YYYY-MM-DD", #16
    "Race year", #17
    "Race location" #18
  )
)

# Display the table
knitr::kable(table_summary, caption = "Variable Summary for main_dta")

vis_miss(main_dta, sort_miss = TRUE, warn_large_data = FALSE) +
  labs(
    title = "Analyzing the Structure of Missingness",
    x = NULL,  # Use NULL instead of "" for x-axis label
    y = "Rows"
  ) +
  theme_minimal(base_size = 14) +  # Use a minimal theme with larger base font size
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Center title and bo
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),  # Adjust x-axis text
    axis.title.y = element_text(size = 14),  # Increase y-axis title size
    panel.grid.major = element_line(color = "lightgray"),  # Light gray gridlines
    panel.grid.minor = element_blank()  # Remove minor gridlines
```

```
  ) +
  scale_fill_brewer(palette = "Set4") +  # Change fill colors to a color palette
  guides(fill = "none")  # Remove the legend

# Create a matrix count of missing values by year and location across all columns
missing_count_matrix <- main_dta %>%
  group_by(year, location) %>%
  summarize(
    missing_count = sum(across(everything(), is.na)),
    .groups = 'drop') %>%
  pivot_wider(names_from = location,
              values_from = missing_count,
              values_fill = 0)

# View the resulting matrix
knitr::kable(missing_count_matrix, caption = "Count of Missingness by Year and Location")

main_dta %<>%
  mutate(completion_time = perc_cr * cr)
# Create a scatterplot of age and completion time
main_dta %>%
  ggplot(aes(x = age_yr, y = completion_time, color = as.factor(sex))) +
  geom_point(alpha = 0.7) +
  labs(title = "Scatterplot of Age and Completion Time",
       x = "Age (years)",
       y = "Completion Time (seconds)") +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # Center title and bol
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),  # Adjust x-axis text
    axis.title.y = element_text(),  # Increase y-axis title size
    panel.grid.major = element_line(color = "lightgray"),  # Light gray gridlines
    panel.grid.minor = element_blank()  # Remove minor gridlines
  )

# Introduce a squared term for age
main_dta %<>%
  mutate(age_yr_sq = age_yr^2)

# Fit models
lm_age_simple <-
  lm(completion_time ~ age_yr,
  data = main_dta)
```

```r
lm_age_sq <-
  lm(completion_time ~ age_yr + age_yr_sq,
  data = main_dta)


# Generate predictions for each model
pred_df <- main_dta %>%
  mutate(pred_simple = predict(lm_age_simple),
         pred_sq = predict(lm_age_sq))

# Create the scatterplot with all model predictions overlaid
pred_df %>%
  ggplot(aes(x = age_yr, y = completion_time)) +
  geom_point(alpha = 0.5) +  # Add points for the actual data
  geom_line(aes(y = pred_simple), color = "blue", linetype = "solid", size = 1) +  # Linear r
  geom_line(aes(y = pred_sq), color = "red", linetype = "solid", size = 1) +  # Quadratic mod
  labs(title = "Scatterplot of Age and Completion Time with Model Fits",
       x = "Age (years)",
       y = "Completion Time (seconds)") +
  scale_color_manual(name = "Model Type",
                     values = c("Linear" = "blue", "Quadratic" = "red")) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # Center title and bol
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),  # Adjust x-axis text
    axis.title.y = element_text(),
    panel.grid.major = element_line(color = "lightgray"),  # Light gray gridlines
    panel.grid.minor = element_blank()  # Remove minor gridlines
  ) +
  theme(legend.position = "right")

# Use modelsummary to display the models
modelsummary::modelsummary(list(
  "Linear Model (Simple)" = lm_age_simple,
  "Linear Model (Quadratic)" = lm_age_sq
  )
)

# Create a correlation matrix of weather measures
weather_vars <- c("tw_c", "perc_rh", "tg_c", "sr_wm2", "dp", "wind", "wbgt", "aqi_median")

# Create summary function for grouped data
summarize_weather_by_location <- function(df, var) {
```

```
  df %>%
    group_by(location) %>%
    summarize(
      n = sum(!is.na(.data[[var]])),
      mean = round(mean(.data[[var]], na.rm = TRUE), 0),
      sd = round(sd(.data[[var]], na.rm = TRUE), 0),
      min = round(min(.data[[var]], na.rm = TRUE), 0),
      max = round(max(.data[[var]], na.rm = TRUE), 0),
      range = round(max(.data[[var]], na.rm = TRUE) - min(.data[[var]], na.rm = TRUE), 0)
    ) %>%
    mutate(variable = var)
}

# Apply function to all weather variables and bind the results
summary_by_location <- bind_rows(lapply(weather_vars, function(var) {
  summarize_weather_by_location(main_dta, var)
}))

# Rearrange columns and arrange by location
summary_by_location <- summary_by_location %>%
  select(location, variable, n, mean, sd, min, max, range) %>%
  arrange(location)

# Create kable table with alternating row colors
knitr::kable(summary_by_location,
      booktabs = TRUE,
      caption = "Summary of Weather Variables by Location")

# Calculate the correlation matrix
cor_matrix <- cor(main_dta[weather_vars], use = "complete.obs")

# Create a correlation matrix plot
corrplot(cor_matrix,
         method = "circle", # or "number", "shade", etc.
         type = "upper", # only show upper triangle
         tl.col = "black", # text color
         tl.srt = 45, # text rotation
         addCoef.col = "black", # add correlation coefficients
         diag = FALSE) # hide the diagonal
# Create performance quartiles
main_dta %<>%
  mutate(completion_time_q = ntile(completion_time, 4))
```

```r
main_dta$completion_time_q <- as.factor(main_dta$completion_time_q)

# Create a boxplot of completion time by flag and performance quartile
main_dta %>%
  ggplot(aes(x = flag, y = completion_time, fill = completion_time_q)) +
  geom_boxplot() +
  labs(title = "Boxplot of Completion Time by Flag and Performance Quartile",
       x = "Flag",
       y = "Completion Time [seconds]",
  fill = "Performance Quartile \n(1) Top 25% \n(2) Top 25-50% \n(3) Top 50-75% \n(4) Bottom 2
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # Center title and bol
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),  # Adjust x-axis text
    axis.title.y = element_text(),
    panel.grid.major = element_line(color = "lightgray"),  # Light gray gridlines
    panel.grid.minor = element_blank()  # Remove minor gridlines
  ) +
  theme(legend.position = "right")

# Create a list to store plots
plot_list <- list()

# Create scatterplots and store them in plot_list
for (var in weather_vars) {
  p <- ggplot(main_dta, aes_string(x = var, y = "completion_time")) +
    geom_point(alpha = 0.1) +
    # Add regression lines for each flag-specific model
    geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE) +
    labs(x = var, y = "Completion Time") +
    theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # Center title and bol
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),  # Adjust x-axis text
    axis.title.y = element_text(),
    panel.grid.major = element_line(color = "lightgray"),  # Light gray gridlines
    panel.grid.minor = element_blank()  # Remove minor gridlines
    )

  plot_list[[var]] <- p
}

# Arrange the plots in a 3x3 grid
do.call(grid.arrange, c(plot_list, ncol = 3))
```

```r
# Fit a regression model
lm_weather <-
  lm(completion_time ~ tw_c + perc_rh + tg_c + sr_wm2 + dp + wind + wbgt + aqi_median,
  data = main_dta)

# Use modelsummary to display the model
modelsummary::modelsummary(lm_weather)
summary(lm_weather)
```