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

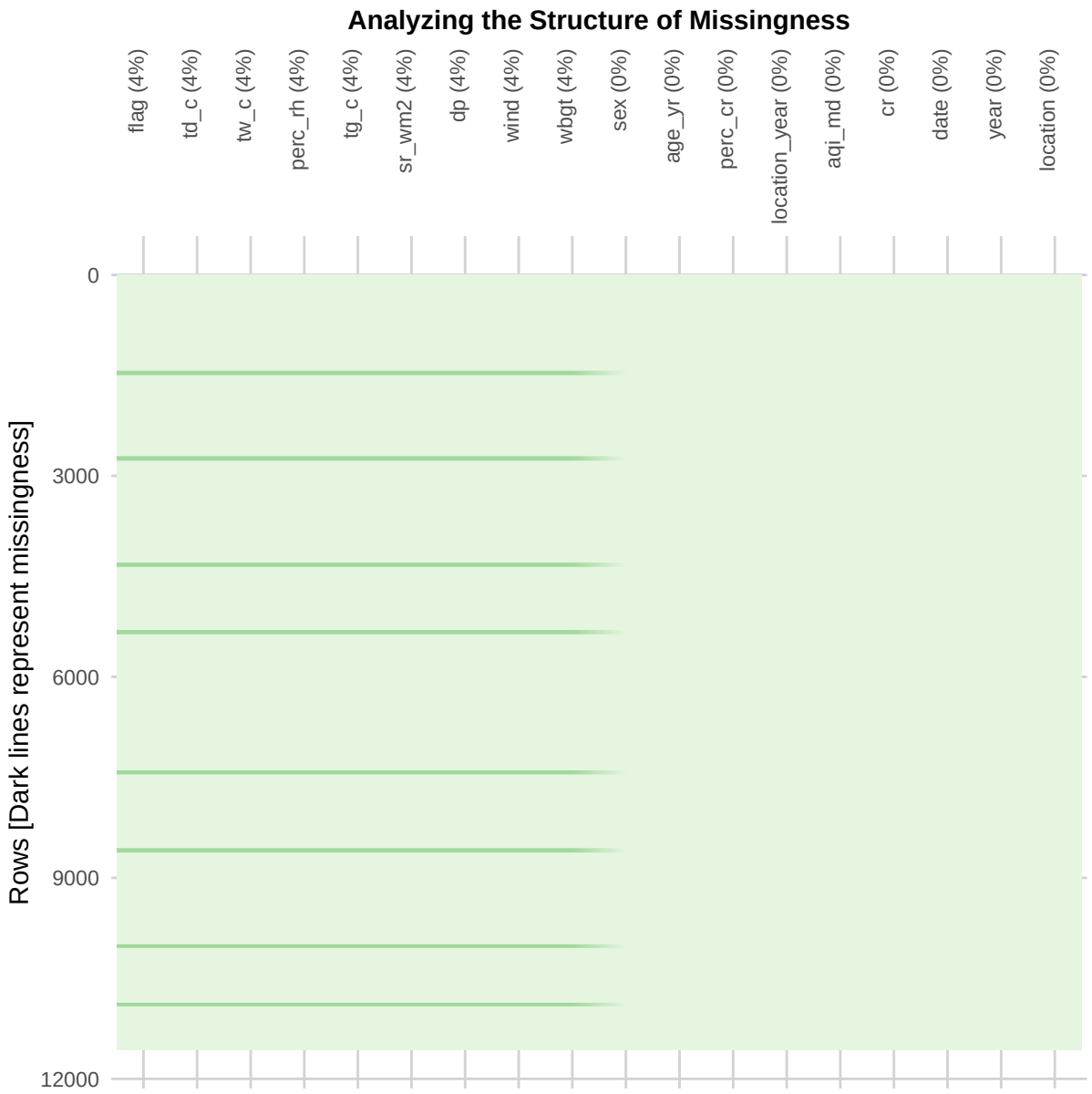
First, let us combine the data from the different datasets into a single dataset. Since each city has one marathon per year, we can create a unique identifier for each marathon by combining the city and the year. We will then use the "project1.csv' dataset, containing the individual-level data, and merge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

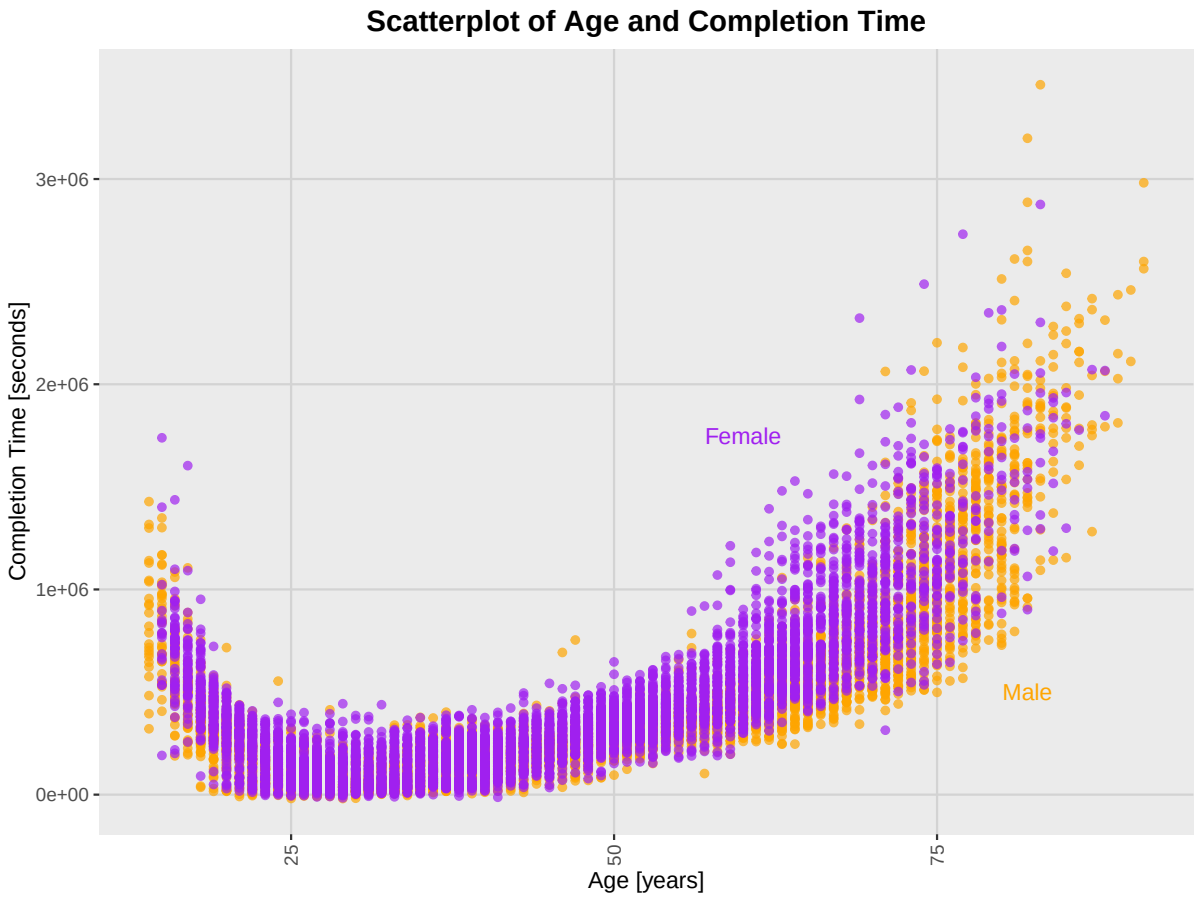
Table 2: Count of Missingness by Year and Location

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, according to 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

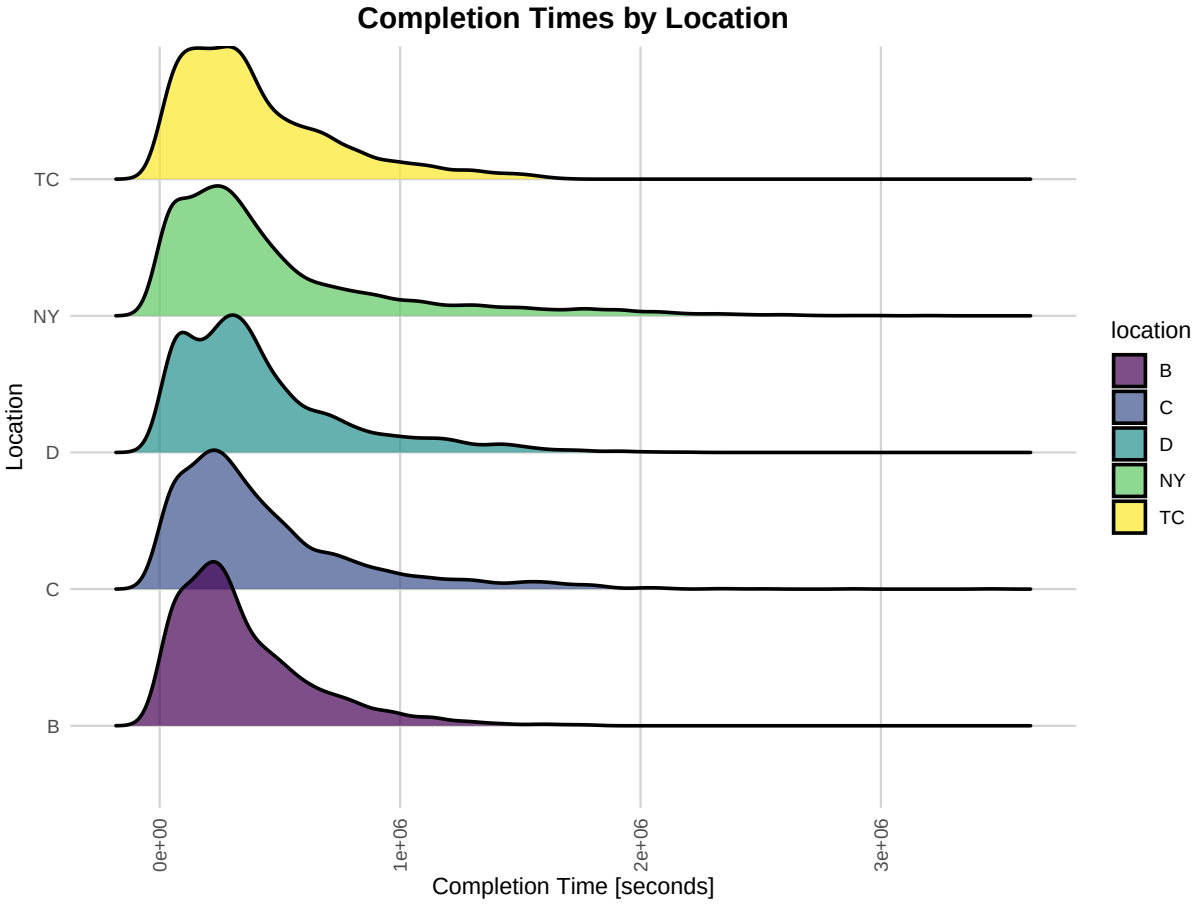
Next, we are interested in the impact of environmental conditions on marathon performance.

Table 3: Weather Variables Summary by Location

	B	C	NY	TC	D
tw_c	7.6 (3.8)	8.5 (5.7)	7.6 (5.0)	9.9 (5.4)	14.9 (2.5)
perc_rh	36.1 (34.2)	60.4 (10.5)	26.9 (30.4)	41.8 (34.2)	49.3 (34.3)
tg_c	24.2 (8.4)	24.5 (6.3)	21.4 (5.9)	24.9 (6.5)	31.6 (7.9)
sr_wm2	649.8 (186.9)	460.5 (94.6)	401.1 (130.9)	435.9 (138.9)	676.8 (190.5)
dp	3.3 (4.5)	4.6 (6.9)	2.7 (7.0)	6.0 (7.3)	12.4 (3.2)
wind	12.0 (4.5)	8.2 (3.2)	11.2 (4.6)	8.8 (3.2)	9.2 (2.9)
wbgt	11.3 (4.5)	12.1 (5.8)	10.7 (4.9)	13.2 (5.4)	18.6 (3.2)
aqi_md	3.0 (3.3)	3.6 (5.6)	3.9 (7.0)	6.5 (4.9)	2.5 (2.9)

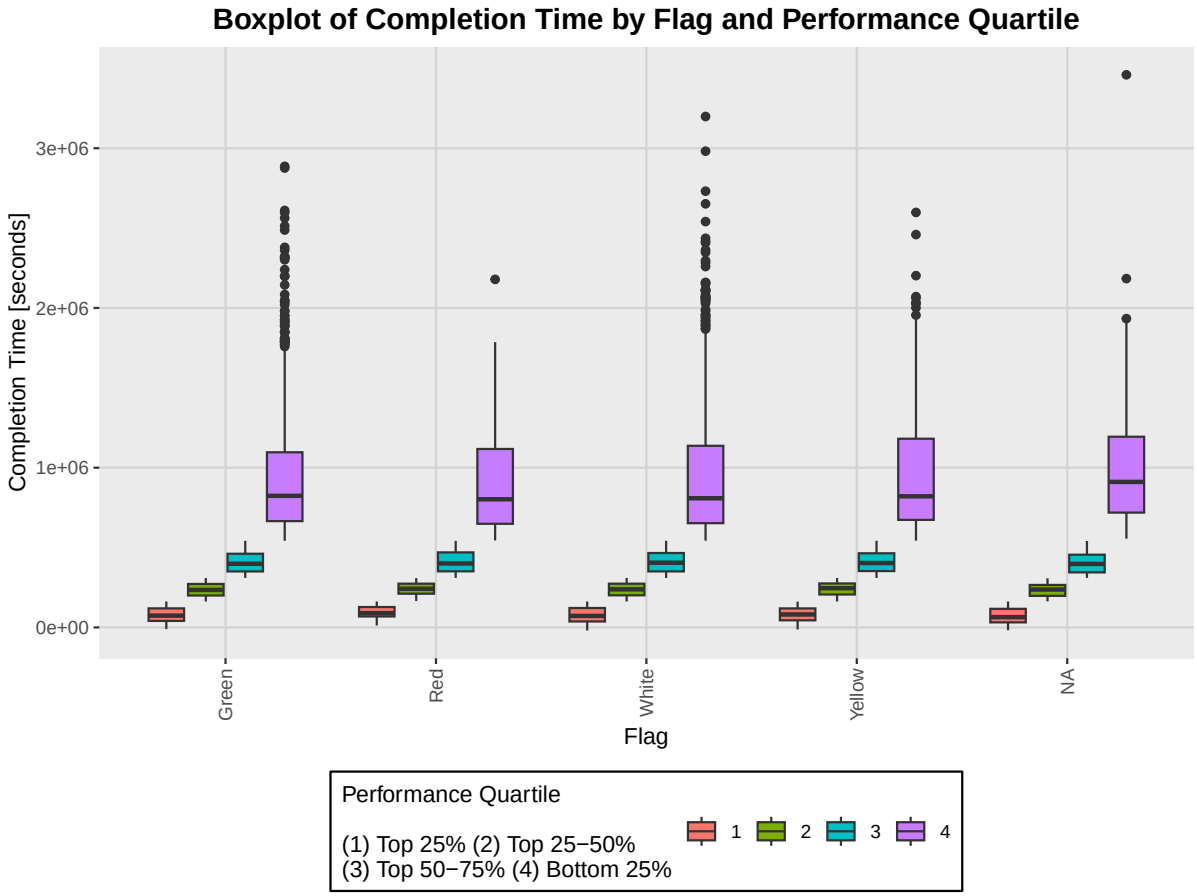
The exploratory analysis shows that the “Grandmas” Duluth marathon has a slightly higher temperature mean than the other marathons. The “Grandmas” Duluth marathon also has the highest solar radiation and dew point mean. With the “Grandmas” Duluth marathon being

the warmest, it would be interesting to see whether completion times are uniformly higher. To visually examine this hypothesis, we will create a ridge plot of completion times by location.



Interestingly, the ridge plot shows that there are no significant differences in completion times across locations. This suggests that the differences in weather conditions across locations do not have a significant impact on baseline completion times.

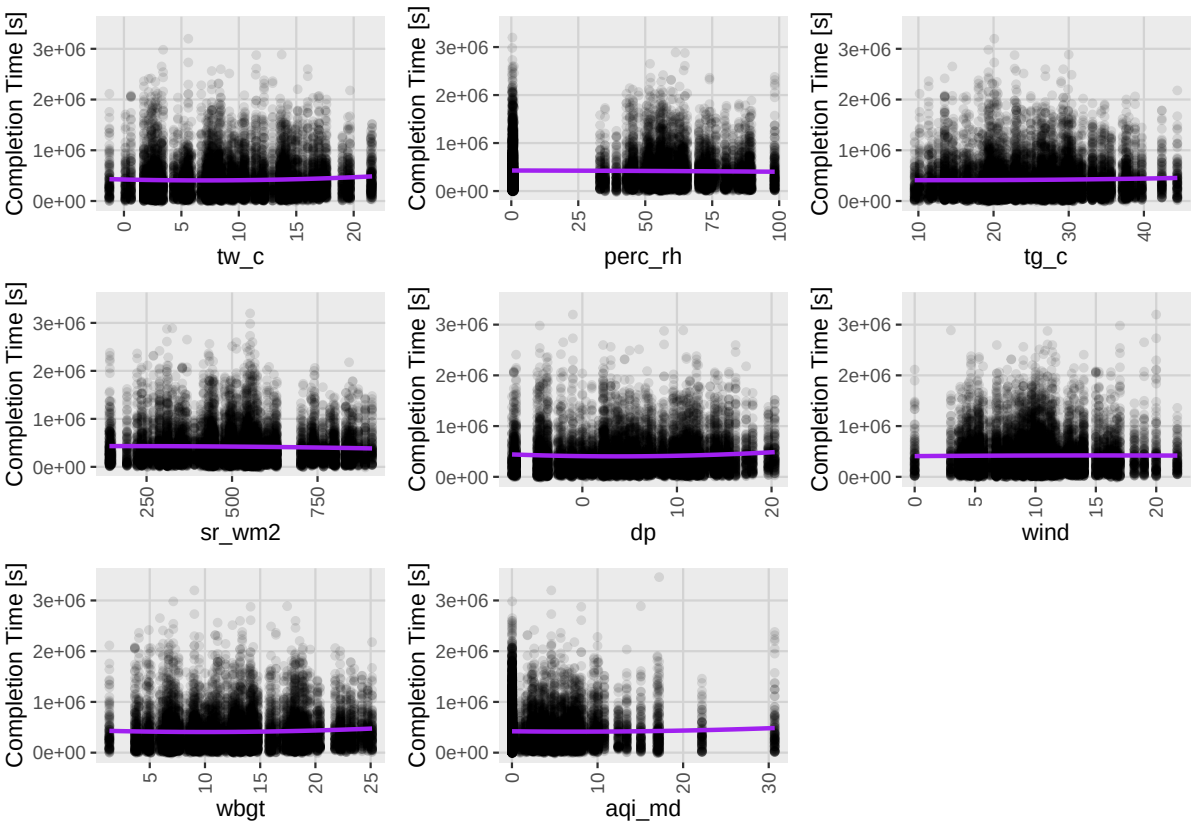
Although there is little evidence that completion times are affected uniformly by temperature variables, we can see slight bimodal trends in the ridge plot in the "Grandmas" Duluth marathon. Potentially, this could be evidence for weather affecting runners differently by performance group. To test whether the relationship between weather and performance varies by performance group, we classify runners into performance quartiles and examine the relationship between completion time and `flag`. The `flag` variable is based on the Wet Bulb Globe Temperature (WBGT) and the risk of heat illness.



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 polynomial line of best fit.

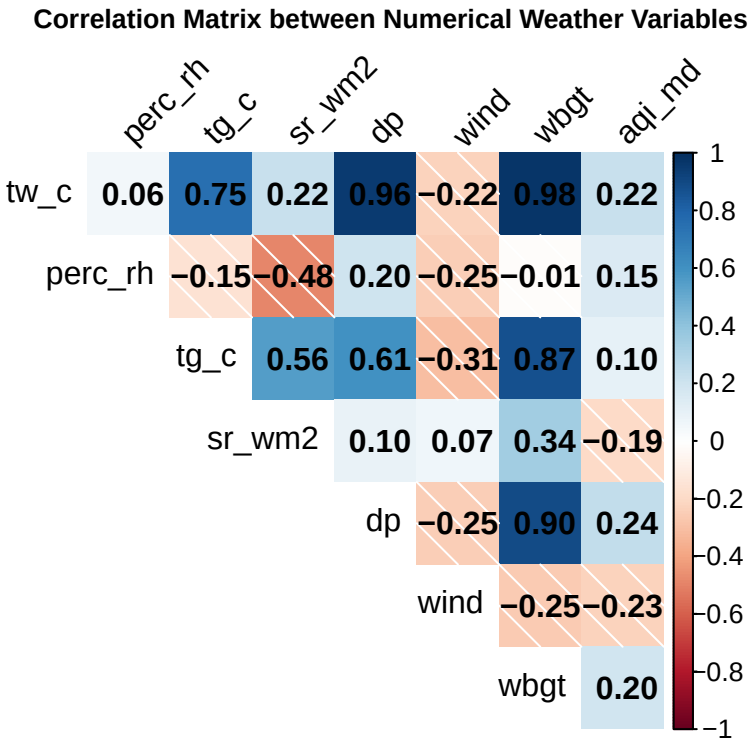


The scatterplots reveal no significant relationships. A lot of the scatterplots show a flat line of best fit and bunching of x-values around discrete values like 0, suggesting that the weather variables do not have a strong impact on completion time.

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. In order to select variables meaningfully, we have to take correlations into account. 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.

Table 4: Summary of the Weather Model

Term	Model Output		
	Coefficient	Standard Error	P-Value
(Intercept)	467255.6134	20694.25381	0.0000000
perc_rh	-627.6614	134.91841	0.0000033
sr_wm2	-165.7725	24.69692	0.0000000
wind	1027.6611	967.76425	0.2883079
wbgt	4374.6545	751.11978	0.0000000
aqi_md	-861.3609	744.88635	0.2475566



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. Finally, 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. One group of variables, i.e. dew point, black globe temperature, wet bulb temperature, and wet bulb globe temperature, are particularly correlated. Thus, we will only include wet bulb globe temperature in the regression analysis, since it contains information from the other variables.

There are some issues of statistical significance, i.e. `wind` and `aqi_md` are not statistically significant at the 10% level. This suggests that there is no meaningful associations between those variables and performance time. The other variables are statistically significant even at the 1% level. This suggests that there is strong evidence that `perc_rh`, `sr_wm2`, and `wbgt` are strongly predictive of completion time. The coefficients on `perc_rh` and `sr_wm2` are negative, suggesting that higher humidity and solar radiation are associated with longer completion times. It is suprising that the coefficient on `wbgt` is positive, suggesting that higher wet bulb globe temperatures are associated with longer completion times. This serves as an important reminder that these relationships may not be evaluated as causal effects, but rather as associations. Future work could explore the positive relationship between wet bulb globe temperature and completion time further and potentially make causal conclusions.

It is difficult to definitively say which factors are "most" significant on grounds of issues regarding statistical estimation and comparing across different units but we can say that solar radiation, relative humidity, and wet bulb globe temperature are the most significant predictors of completion time.

References

- Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. Med Sci Sports Exerc, 42(1), 135-41.
- Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. Medicine and science in sports and exercise, 39(3), 487-493.
- Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.
- Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., … & Millet, G. Y. (2022). Sex differences in endurance running. Sports medicine, 52(6), 1235-1257.
- Yanovich, R., Ketko, I., & Charkoudian, N. (2020). Sex differences in human thermoregulation: relevance for 2020 and beyond. Physiology, 35(3), 177-184.

Code Appendix

```
# 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(broom)
library(corrplot)
library(kableExtra)
library(grid)
library(gridExtra)
library(ggridges)
library(knitr)
library(ggplot2)
library(naniar)
library(gtsummary)

# 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
```

```r
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) %>%
```

```r
  mutate(location_year = paste0(location, "_", year))

# Clean aqi_values
aqi_trim <- aqi_values %>%
  select(location_year, arithmetic_mean) %>%
  group_by(location_year) %>%
  summarize(
    aqi_md = 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"))

# 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 = Male)", #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 [Dark lines represent missingness]"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
    axis.title.y = element_text(size = 14),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank()
  ) +
  scale_fill_brewer(palette = "Set4") +
  guides(fill = "none")

# Create a matrix count of missing values by year and location
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)
```

```r
# 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"),
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
    axis.title.y = element_text(),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank()
  ) +
  scale_color_manual(values = c("purple", "orange")) +
  guides(color = "none") +  # Remove legend
  annotate("text", x = 60, y = 1750000, label = "Female", color = "purple") +
  annotate("text", x = 82, y = 500000, label = "Male", color = "orange")

# 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)

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, color = "grey") +
  geom_line(aes(y = pred_simple), color = "orange",
                linetype = "solid", size = 1) +  # Linear model
  geom_line(aes(y = pred_sq), color = "purple",
                linetype = "solid", size = 1) +  # Quadratic model
  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 bold
    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"),
    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_md")

locations <- unique(main_dta$location)

# Create a summary function that formats mean and sd
summary_fn <- function(data, var) {
```

```r
  mean_val <- mean(data[[var]], na.rm = TRUE)
  sd_val <- sd(data[[var]], na.rm = TRUE)
  return(sprintf("%.1f (%.1f)", mean_val, sd_val))  # Rounded to 1 digit
}

# Generate the summary table
summary_table <- sapply(locations, function(loc) {
  sapply(weather_vars, function(var) {
    summary_fn(main_dta %>% filter(location == loc), var)
  })
})

# Convert to a data frame and set row names
summary_df <- as.data.frame(summary_table, row.names = weather_vars)

# Create a kable table
kable(summary_df, caption = "Weather Variables Summary by Location") %>%
  kable_styling(full_width = F)

# Ridge plot
ggplot(main_dta, aes(x = completion_time, y = location, fill = location)) +
  geom_density_ridges(alpha = 0.7, size = 0.8, scale = 1.2) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
    axis.title.y = element_text(),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank()
  ) +
  labs(title = "Completion Times by Location",
       x = "Completion Time [seconds]",
       y = "Location") +
  scale_fill_viridis_d()  # Adjust color scale for the fill
# Create performance quartiles
main_dta %<>%
  mutate(completion_time_q = ntile(completion_time, 4))

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

# Box plot
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% (2) Top 25-50% \n(3) Top 50-75% (4) Bottom 25%") +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
    axis.title.y = element_text(),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank(),
    legend.position = "bottom",
    legend.box = "vertical",
    legend.background = element_rect(fill = "white", color = "black"),
    legend.key = element_rect(fill = "white")
  )

# 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, color = "purple") +
    labs(x = var, y = "Completion Time [s]") +
    theme(
      plot.title = element_text(hjust = 0.5,
      size = 14, face = "bold"),  # Center title and bold
      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"),
      panel.grid.minor = element_blank()
  )

  plot_list[[var]] <- p
}
```

```r
# Arrange the plots in a 3x3 grid with a global title
grid.arrange(grobs = plot_list, ncol = 3,
  top = grid::textGrob("Weather Variables by Completion Time \n",
                       gp = gpar(fontsize = 14, fontface = "bold")))
# Calculate the correlation matrix
cor_matrix <- cor(main_dta[weather_vars], use = "complete.obs")

# Create a correlation matrix plot
corrplot(cor_matrix,
         method = "shade", # or "number", "shade", etc.
         type = "upper", # only show upper triangle
         mar=c(0,0,2,0),
         tl.col = "black", # text color
         tl.srt = 45, # text rotation
         addCoef.col = "black", # add correlation coefficients
         diag = FALSE, # hide the diagonal
         title = "Correlation Matrix between Numerical Weather Variables",
         cex.main = 0.8)

# Fit a regression model
lm_weather <-
  lm(completion_time ~ perc_rh + sr_wm2 + wind + wbgt + aqi_md,
  data = main_dta)

# Extract the relevant statistics from the lm_weather model
lm_weather_tidy <- tidy(lm_weather)

# Create a kable table with coefficient, p-value, and standard error
lm_weather_tidy %>%
  select(term, estimate, std.error, p.value) %>%
  kbl(col.names = c("Term", "Coefficient", "Standard Error", "P-Value"),
      caption = "Summary of the Weather Model") %>%
  kable_classic(full_width = F, html_font = "Cambria") %>%
  add_header_above(c(" " = 1, "Model Output" = 3)) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```