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:

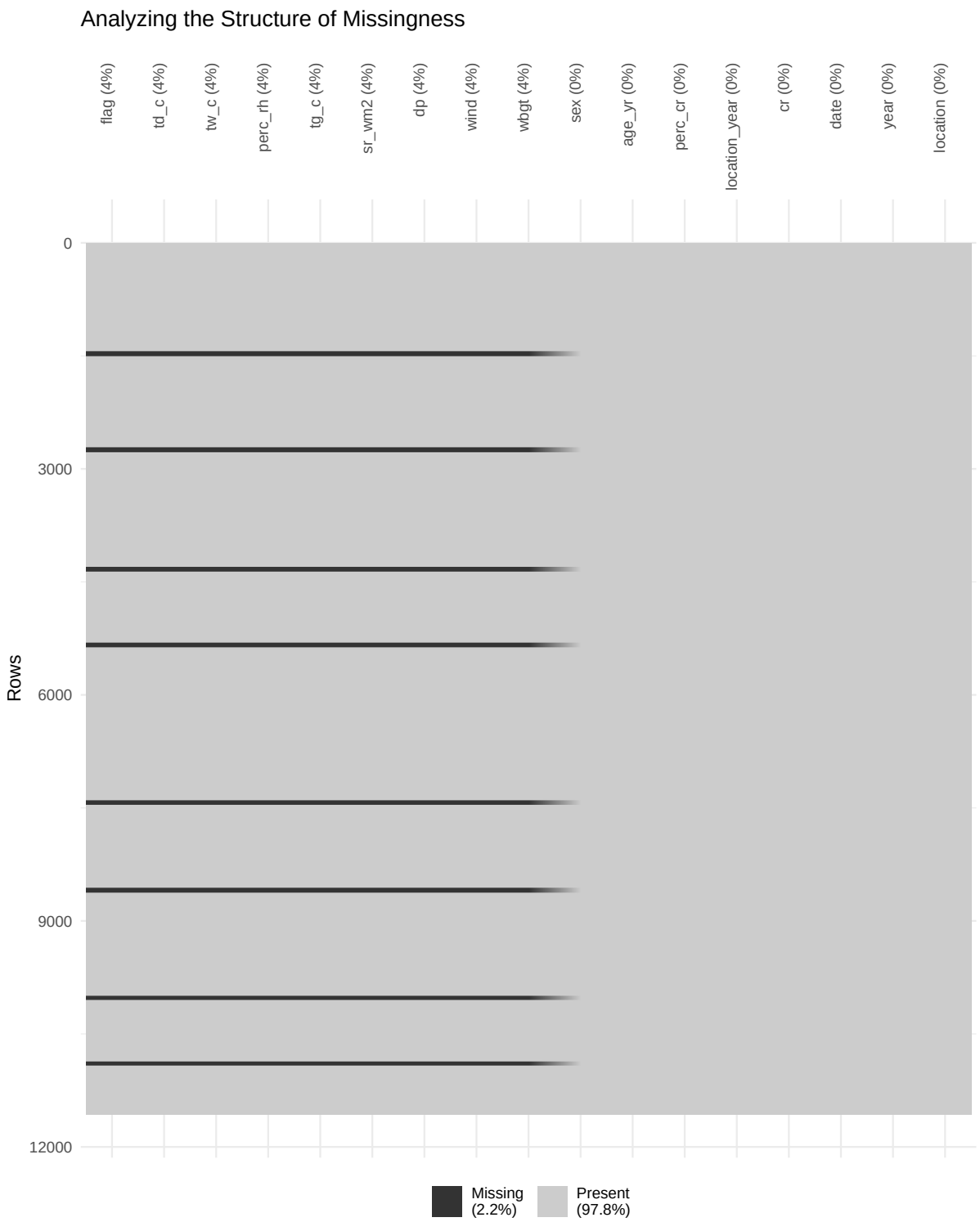
- Aim 1: Examine effects of increasing age on marathon performance in men and women
- Aim 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.
- Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc)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

```
tibble [11,564 x 17] (S3: tbl_df/tbl/data.frame)
 $ sex          : num [1:11564] 1 1 1 1 1 1 1 1 1 1 ...
 $ flag         : Factor w/ 4 levels "Green","Red",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ age_yr       : num [1:11564] 18 19 20 21 22 23 24 25 26 27 ...
 $ perc_cr      : num [1:11564] 35.7 39.3 15.7 7.9 24.7 ...
 $ td_c         : num [1:11564] 13.8 13.8 13.8 13.8 13.8 ...
 $ tw_c         : num [1:11564] 8.23 8.23 8.23 8.23 8.23 ...
 $ perc_rh      : num [1:11564] 45.6 45.6 45.6 45.6 45.6 ...
 $ tg_c         : num [1:11564] 28 28 28 28 28 ...
 $ sr_wm2       : num [1:11564] 766 766 766 766 766 ...
 $ dp           : num [1:11564] 2.23 2.23 2.23 2.23 2.23 ...
 $ wind         : num [1:11564] 12.7 12.7 12.7 12.7 12.7 ...
 $ wbgt         : num [1:11564] 12.7 12.7 12.7 12.7 12.7 ...
 $ location_year: chr [1:11564] "B_2016" "B_2016" "B_2016" "B_2016" ...
 $ cr           : num [1:11564] 8337 8337 8337 8337 8337 ...
 $ date         : Date[1:11564], format: "2016-04-18" "2016-04-18" ...
 $ year         : num [1:11564] 2016 2016 2016 2016 2016 ...
 $ location     : chr [1:11564] "B" "B" "B" "B" ...
```

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 (~ 4%). This suggests that the missingness is not random, but rather systematic.

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 will first outline an analysis strategy:

- Variable of interest: `completion_time` = `perc_cr` · `cr` (i.e. "percentage off course record" times "course record")

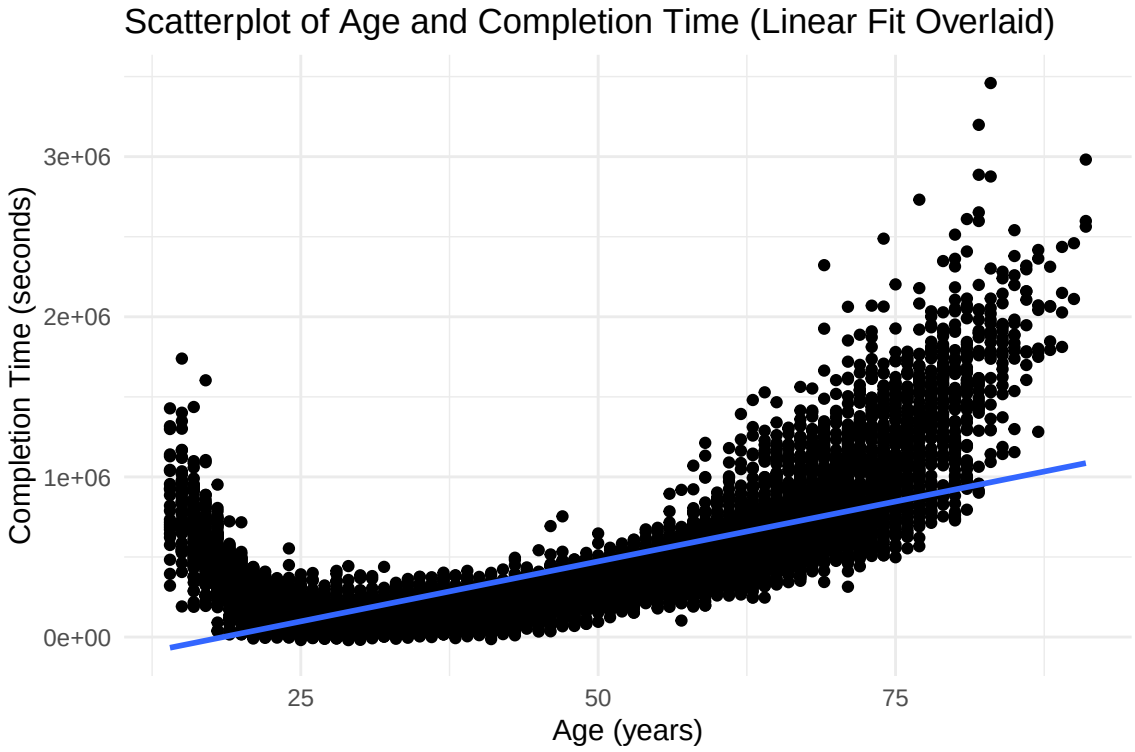
- Predictor: `age_yr`
- Individual-level controls: `sex`, `flag`
- Race-level controls: `flag`, `wind`

where `flag` is constrcuted from both individual-level and race-level information. We will use `flag` as a proxy for weather conditions.

Moreover, we will test between two modeling frameworks:

- Modeling a linear trend, i.e. ordinary least squares regression
- Modeling a non-linear trend, i.e. introduction of a non-linear term or logaritmic transformation

Naturally, 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. However, exploratory analysis suggests that the relationship between age and completion is not overly complex, albeit being non-linear. Moreover, the data are relatively lower-dimensional, making the fitting of a ordinary least squares regression model feasible. The risk of overfitting is relatively low when compared to using more complex models.



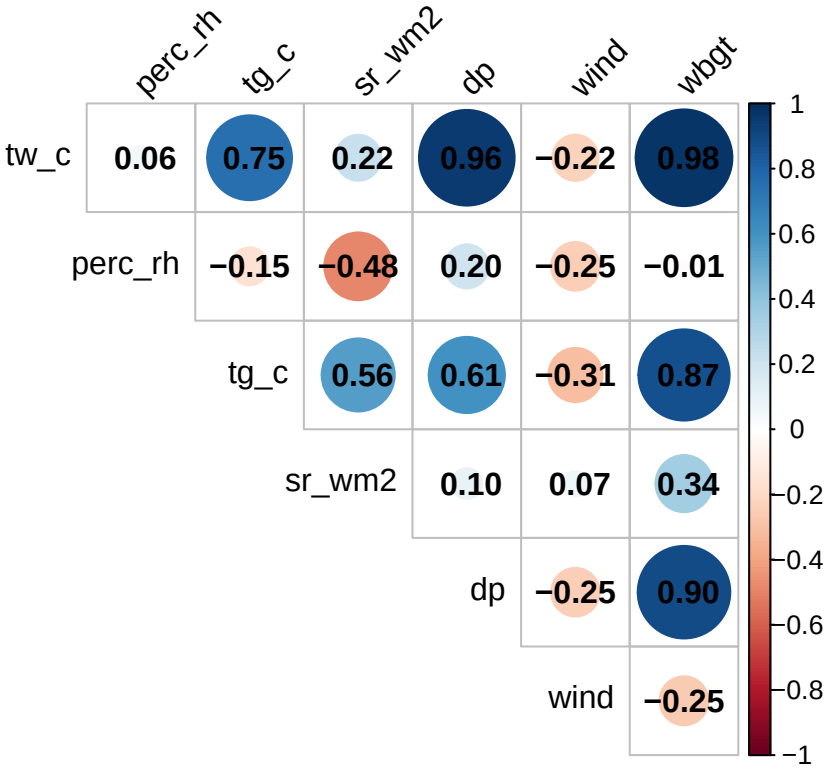
It is straightforward to see that there is a non-linear relationship between age and completion time. Until roughly age 25, completion time decreases. From age 25 to age 35, completion

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.

In terms of model selection, it is clear that a linear model will not be able to capture the non-linear relationship between age and completion time. However, a linear regression model may still be appropriate if we introduce a squared term for age.

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

An initial concern surrounding variable selection is the possibility of two or more variables within our model being highly correlated. Ideally, we want to avoid that by choosing variables that are important but not repititive. When looking at the data, we have access to multiple measures of temperature and weather. Let us turn to those first.



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

	Linear Model (Simple)	Linear Model (Log-Log)	Linear Model (Quadratic)
(Intercept)	−265 309.672	7.340	1 070 564.738
	(<0.001)	(<0.001)	(<0.001)
age_yr	15 075.569		−49 822.607
	(<0.001)		(<0.001)
sex	−40 743.839	−0.167	−81 393.063
	(<0.001)	(<0.001)	(<0.001)
flagRed	68 897.260	0.289	75 104.831
	(<0.001)	(<0.001)	(<0.001)
flagWhite	−21 452.933	−0.103	−20 653.617
	(<0.001)	(<0.001)	(<0.001)
flagYellow	37 839.176	0.100	39 544.819
	(<0.001)	(<0.001)	(<0.001)
wind	203.186	0.004	−220.791
	(0.769)	(0.079)	(0.608)
log(age_yr)		1.391	
		(<0.001)	
age_yr_sq			686.816
			(<0.001)
Num.Obs.	11 073	11 039	11 073
R2	0.495	0.329	0.805
R2 Adj.	0.495	0.329	0.805
AIC	308 629.0		298 112.8
BIC	308 687.5		298 178.6
Log.Lik.	−154 306.481	−14 083.180	−149 047.403
F	1809.486		6515.724
RMSE	272 786.05	0.87	169 650.72

as WGBT is derived from temperature, for example. It serves as a warning sign for the analyst, however, as they attempt to avoid estimation issues on ground of collinearity.

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

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(corrplot)
library(kableExtra)
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"))
```

```r
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))
```

```r
# Merge data datasets with project1_dta
main_dta <- project1_dta %>%
  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"))

kable(str(main_dta))

vis_miss(main_dta, sort_miss = TRUE, warn_large_data = FALSE) +
labs(title = "Analyzing the Structure of Missingness") +
  xlab("") +
  ylab("Rows") + theme(axis.text.x = element_text(angle = 90, hjust = 1))
# 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
kable(missing_count_matrix)

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)) +
```

```
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatterplot of Age and Completion Time (Linear Fit Overlaid)",
       x = "Age (years)",
       y = "Completion Time (seconds)") +
  theme_minimal()
# 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 + sex + flag + wind,
  data = main_dta)

lm_age_sq <-
  lm(completion_time ~ age_yr + age_yr_sq + sex + flag + wind,
  data = main_dta)

lm_age_ln <-
  lm(log(completion_time) ~ log(age_yr) + sex + flag + wind,
  data = main_dta)

# Use modelsummary to display the models
modelsummary::modelsummary(list(
  "Linear Model (Simple)" = lm_age_simple,
  "Linear Model (Log-Log)" = lm_age_ln,
  "Linear Model (Quadratic)" = lm_age_sq
  ),
  statistic = "p.value"
)

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

# 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
```