Andrea Kuhn & Alex Nguyen
Stat 6950

Data Project

About the Data

The data set contains results for the marathon, 50 km race-walk, 20 km race-walk, 10000 m, 5000 m, and 3000 m-steeplechase for major competitions (Commonwealth Games, Diamond League, World Athletics Continental Cup, World Athletics Gold Label Races, Olympic Games, World Athletics Race Walking Team Championships, and World Championships) from 2000-2019. The standing event record and world record were also included for each event. These results were obtained from the official websites and collected between February 2016 and March 2024.

The original data set contained 1258 races ranging from 1936 to 2019 after screening out races at the beginning of the creation due to inconsistencies in community interest and runner participation. For our uses, we will only be considering race-walking, track, and marathon championships races from 2000-2010 and marathon Gold Label Races from 2010-2019. Additionally, since our focus will be only on the top three results (elite athletes) rather than the 25th, 50th, 100th, and 300th place finishers (well-trained runners) we removed several Gold Label Races (marathons) that are historically not attended by elite athletes (Athens, Gold Coast, Ottawa, Philadelphia, etc.). These races were selected by reviewing the results from 2010-2019 and by utilizing personal knowledge and experience. This choice to reduce the original data set was also driven by the necessity to manually compete the data for the remaining Gold Label Races. Because most Gold Label Races (marathons) have a combined male and female start, the original data set did not distinguish between the genders for these races. Therefore, we completed our data set by copying the weather and location values from the male race and then adding the top three female finishers. We also removed races that were missing finishing times from the original data set. Our current data set has 657 races ranging from 2000-2019.

For the process used to collect the weather data, see Mantzios et al..

Proposed Question

Our question of interest will be how weather variables at the time of the race impact percent deviation of the finishing times for elite athletes from either the standing event record or standing world record. The goal is to explore what weather variables impact performance the greatest. Further, we will potentially break down this relationship across several categories such as gender, race distance, and competition type (because championship style races are typically ran slower due to the focus on placing rather than finishing time).

Another component we will assess is whether there has been a shift in performance since 2017. We believe 2017 may carry special significance, especially for road races, due to the introduction of "super shoes" (i.e. Nike Vaporflys).

References

Mantzios, Konstantinos et al. "Effects of Weather Parameters on Endurance Running Performance: Discipline-specific Analysis of 1258 Races." Medicine and science in sports and exercise vol. 54,1 (2022): 153-161. doi:10.1249/MSS.0000000000002769

Data Cleaning

```r
results = read.csv("race_results.csv", header = TRUE)

## rename columns for easier reading
colnames(results) = c("competition", "distance", "sex", "host", "country",
                      "day", "month", "year", "time",
                      "latitude", "longitude", "NOAA_ID", "station_loc",
                      "dist_from_loc", "air_temp", "dew_point", "wind_speed",
                      "adj_wind_speed", "rel_hum", "clouds_OKTA",
                      "diff_from_req_time", "time_zone", "solar_rad",
                      "heat_index", "simp_wet_bulb", "wet_bulb",
                      "world_record", "standing_record", "time_1st",
                      "time_2nd", "time_3rd", "time_4th", "time_5th",
                      "time_6th", "time_7th", "time_8th", "time_9th",
                      "time_10th", "time_avg")

## remove missing values
missing = which(results$world_record == "" |
                  results$standing_record == "" |
                  results$time_1st == "")
results = results[-missing,]

## convert time string to time object
results$world_record = as_hms(results$world_record)
results$standing_record = as_hms(results$standing_record)
results$time_1st = as_hms(results$time_1st)
```

```r
results$time_avgTop3 = as_hms(results$time_avgTop3)

## convert time to seconds
results$world_record_s = as.numeric(results$world_record)
results$standing_record_s = as.numeric(results$standing_record)
results$time_1st_s = as.numeric(results$time_1st)

## percent deviation from world and standing record
results = results %>%
  mutate(percent_world = (time_1st_s - world_record_s) /
           world_record_s * 100,
         percent_standing = (time_1st_s - standing_record_s) /
           standing_record_s * 100)
```