Andrea Kuhn & Alex Nguyen
Stat 6950

Data Project

Proposal

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 Marathons, 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.

Additionally, the 4th through 10th place finishing times were not provided for a large portion of races, especially the marathon distance, and it was determined to be beyond the scope of this project to collect these results. Therefore, we cannot confidently use these variables or any function of these variables as the response for our model, due to the large number of observations with missing values. However, for all races included in our final data set, the top 3 finishing times were available. Therefore, we are able to use these variables or a function of these variables as the response. We also found 7 races with a missing meteorological station ID and missing weather data. These races were removed from our data set. Our current data set has 650 races ranging from September 2000 to December 2019. For the process used to collect the weather data, see Mantzios et al..

Question of Interest

Our goal is to assess the impact of weather, specifically temperature or heat related variables, on elite athlete performance in endurance running events. We hope to determine which, if any, weather variables influence performance, and whether the gender of participants, competition format, or race length also play a role.

Exploratory Data Analysis

Our data set contains 12 race identifier variables, 16 weather variables, and 19 race result variables with 7 being generated or derived from the original unique 12 race result variables. The race identifier variables include the purpose of the race, distance, and time and location details. Most of the weather variables are numeric and can be considered continuous. Similarly, the race result variables are numeric and can be considered continuous. To simplify calculations with time variables, we transform the records and finishing times to seconds rather than hours, minutes, and seconds.

Additionally, we separated `dist_from_loc` into a categorical variable. While the majority of races are situated within a 10-mile radius of a meteorological station, there are instances where races are located considerably farther away (491 races within 10-miles vs. 159 races further than 10-miles). By categorizing races based on their proximity to the closest meteorological station, we can account for potential inaccuracies in weather data, particularly for events located further from monitoring infrastructure.

Response Variable

Our response variable `percent_standing_1st` represents the percent difference between the first place finishing time and the event record going into the start of the event. A negative value indicates the finisher set a new event record, while a positive value indicates the finisher had a time slower than the event record. The values range from -8.604% to 14.603% with an average of 1.592%. Figure 1 shows the distribution is uni-modal and roughly symmetric with a slight right skew. It does not appear that any transformations of the response variable are necessary.

There appear to be four potential outliers in our response variable with values above 10%. These observations correspond to the 2018 Men's and Women's Boston Marathon and the 50K and 20K race-walk at the 2019 World Championships in Doha, Qatar. The 2018 Boston Marathon experienced severe wind and rain, as well as cold weather that forced many of the top competitors to drop out with hypothermia symptoms. The 2019 World Championship race-walks started just before midnight in Doha in an attempt to avoid the blistering desert heat. However, even at midnight, the high temperature and humidity produced brutal conditions. Therefore, we will not remove these observations as outliers. The results of the four observations with response values below -5% were also confirmed to be correct.

Weather Variables

We expect collinearity between several of the weather predictors as many are calculated using one another. For example, the heat index was calculated using either air temperature and dew point or air temperature and relative humidity; the adjusted wind speed was calculated using wind speed and then adjusting for buildings and elevation; and the wet bulb globe temperature (WBGT) and simple WBGT were calculated using previous methodology that relies on standard meteorological data (air temperature, dew point, etc.) to predict the WBGT without completing a true measurement. Other implicit relationships also occur without direct dependence. For example, when cloud coverage is high, solar radiation will likely be low.

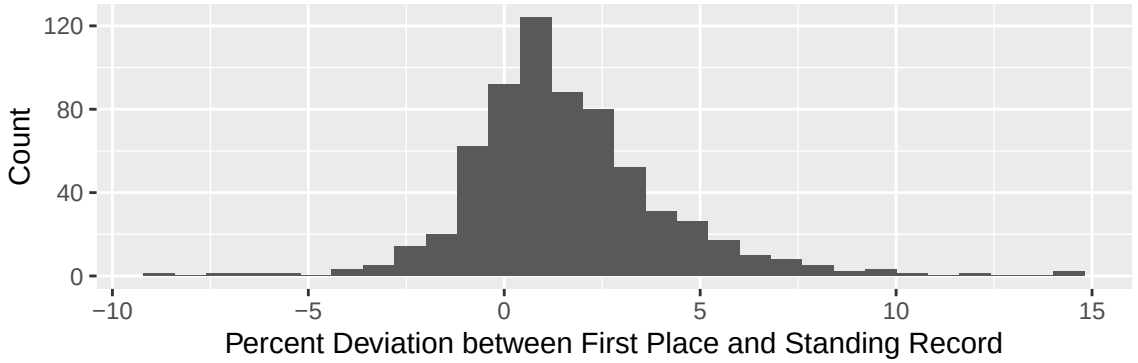
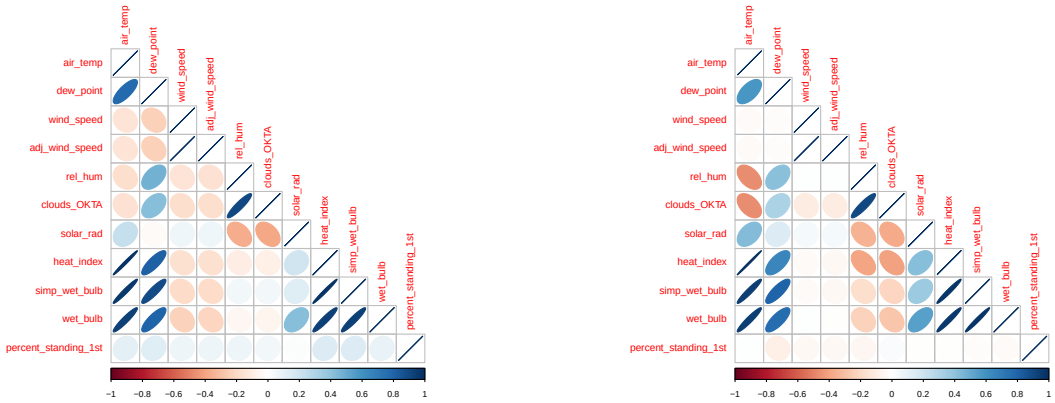


Figure 1: Histogram of response variable.

We may note the differences in the correlation matrix between the weather variables and the percent deviation when races are separated by those within 10-miles of the meteorological station (Figure 2a) and those further than 10-miles (Figure 2b). We observe higher overall correlation between percent deviation and the weather data for races that are within 10 miles of a meteorological station as compared to races further away. Additionally, some relationships between predictors and the response appear to flip direction. This suggests that weather conditions may play a more influential role in races held closer to meteorological stations.



(a) Weather station within 10-miles.

(b) Weather station beyond 10-miles.

Figure 2: Correlation plot between weather predictors and response.

Regardless of proximity, visually assessing scatterplots and calculating correlation coefficients, it appears that air temperature, dew point, wind speed, heat index, and simplified WBGT are likely to be important in predicting the percent deviation between finishing time and event record. The scatterplot matrix with these predictors and all races is shown in Figure 3. However, due to the high collinearity between these predictors, it is also possible that they are explaining the same variation in percent deviation. Formal tests will need to be completed to determine which predictors provide the best model fit and whether proximity to the meteorological station influences the effect of the weather predictors. There does not appear to be any outliers in our weather predictors of interest nor does a transformation of any of the predictors seem necessary.

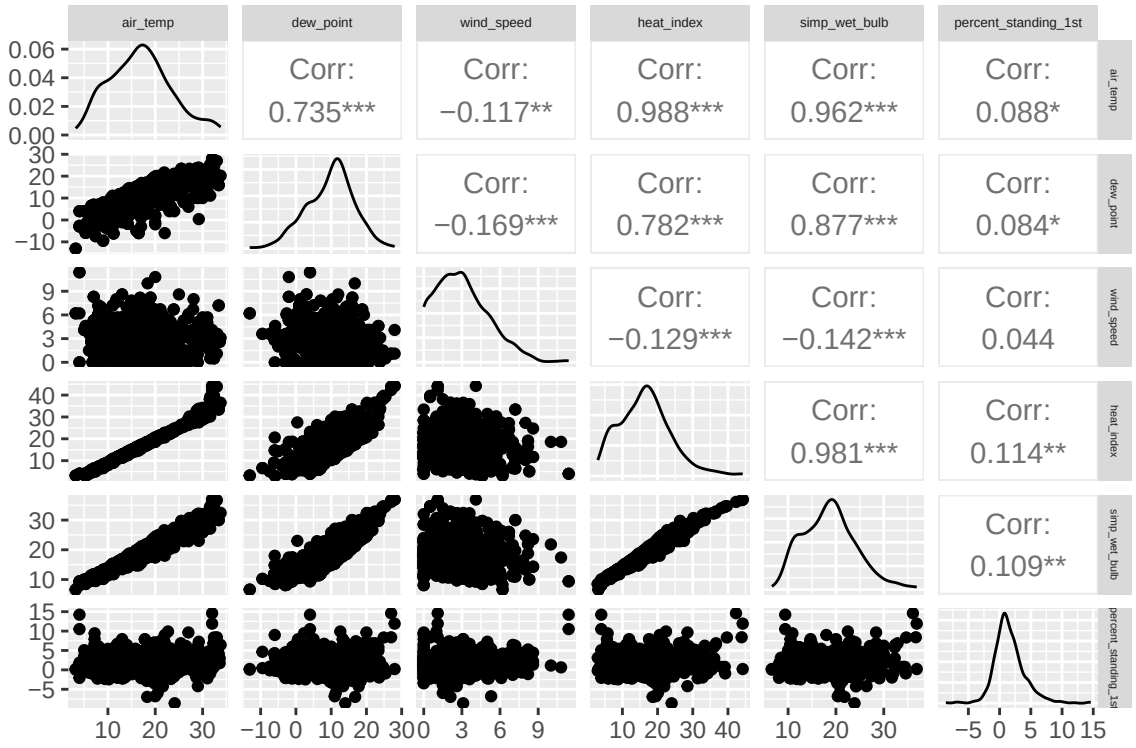


Figure 3: Scatterplot matrix of significant weather predictors.

Categorical Variables

While our primary goal is to assess the impact of weather variables on the performance of elite endurance athletes, several categorical variables display a potential relationship with percent deviation from the event record.

Beginning with analyzing our available categorical variables and their individual distributions, we see that the split between male and female races is roughly even with 357 male races and 293 female races. This difference is largely due to a greater number of Gold Marathons, 50K race-walks, and 3K steeplechase results for males. The 3K steeplechase was not introduced to major female competitions until the late 2000s, and the 50K race-walk had no major female results during our data collection time period. Additionally, we see there is a relationship between competition and distance. For example, the Diamond League only includes results at 5K and 10K, while Gold Marathon only includes results for the marathon. We see these counts in Table 1.

Now, assessing the relationships between our categorical variables and our response variable, we see that the percent deviation from the event record appears to show some variability across competition type (Figure 4). Specifically, the Common Wealth and Olympic races tend to have deviations either closer to zero or more negative. Additionally, Gold Marathons and World Championships races appear to have several right-skewed observations with finishing times significantly slower than the event record.

Figure 5 shows the relationships between date, distance, gender, and percent deviation from the standing record.

Table 1: Count data of primary categorical variables.

	Men						Women					
Comp.	10K	20K	3K	50K	5K	Mar.	10K	20K	3K	50K	5K	Mar.
Common Wealth	5	4	5	2	4	5	4	3	3	0	5	5
Diamond League	0	0	15	0	48	0	0	0	14	0	36	0
Gold Marathon	0	0	0	0	0	166	0	0	0	0	0	146
Olympic	5	2	5	2	4	5	5	2	2	0	5	3
World Champ.	10	9	8	8	9	10	10	8	6	0	8	12
World Cup	0	8	5	9	4	0	0	9	3	0	4	0

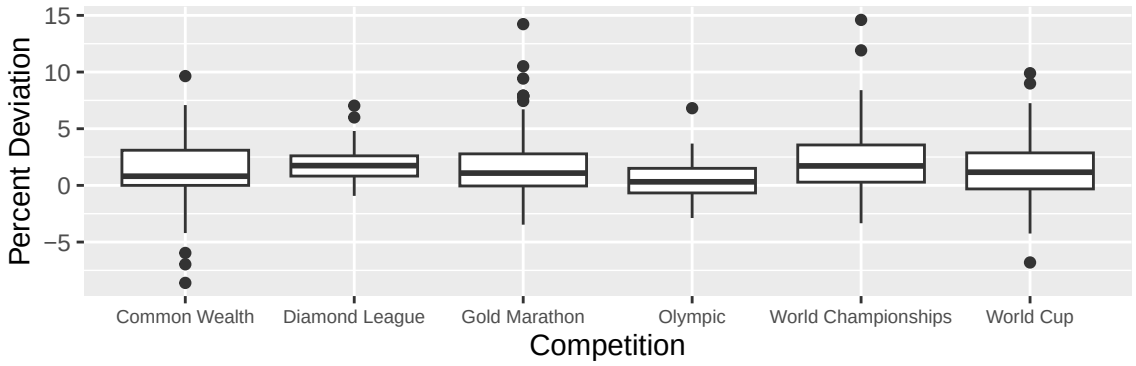


Figure 4: Percent Deviation by Competition Type.

Gender does not appear to have a significant relationship with percent deviation from the event record by itself. However, in addition to possibly accounting for the discrepancy in sample sizes for distances and competitions across genders, including gender as part of an interaction term would allow us to assess the differences in the effects of weather on endurance performance between males and females. Other interaction terms between our categorical variables and weather predictors may also be beneficial for model fit.

In Figure 5, we also see that both genders had a slightly higher proportion of record setting races in the mid 2000s before returning to a more stable rate of records by 2010. Similar to competition type, distance does not appear to display an overwhelmingly strong relationship with percent deviation; however, there may be some relationship. For example, marathons appear to have a higher frequency of relatively slow finishing times. In relation to date, there may possibly be a slight downward trend in percent deviation for women since 2010. Another observation to note is the increase in the number of race results after 2010. This is due to our manual entry of Gold Marathon results for women between 2010 and 2019.

To demonstrate the possibility of an interaction term between a weather predictor and a categorical variable, Figure 6 shows a relationship between heat index and distance. It is possible that the slope of a regression line for `percent_standing_1st` on `heat_index` will differ for each level of `distance`.

What are the key figures or numerical summaries that describe the most important aspects of the data? Likely will be simple correlation between weather predictors and response. Something else?

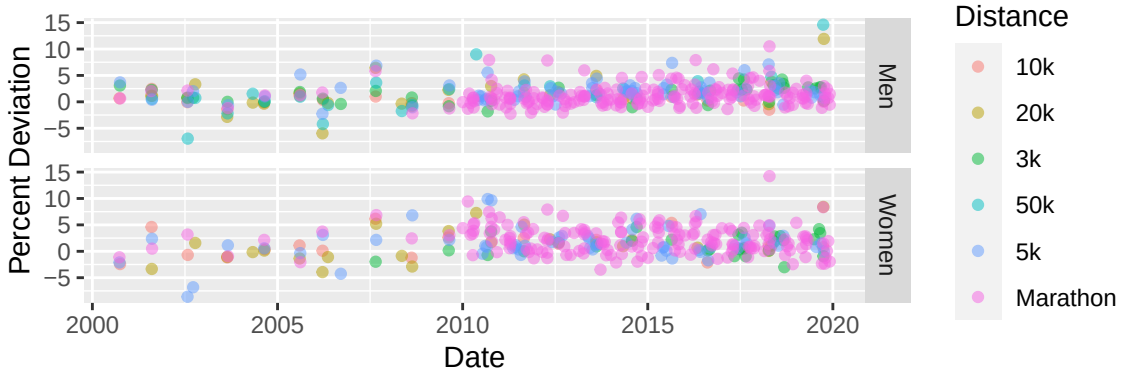


Figure 5: Percent Deviation by Date, Distance, and Gender.

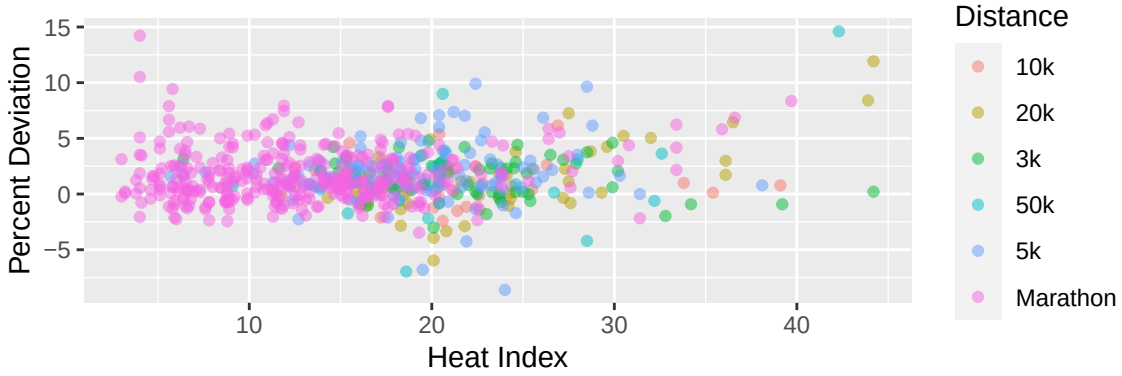


Figure 6: Percent Deviation vs. Heat Index, separated by distance

Does your EDA suggest to you what modeling approaches you should aim to try?
MLR? Polynomial? Likely want to keep model simple to aid in interpretation of final parameters. Goal is to say "a one unit increase in weather is associated with a % drop in performance."
Start with all variables and go from there?

Outline how you propose to model the data to investigate your question of interest in a brief paragraph. You shouldn't necessarily be fitting any models at this stage, but rather exploring the data to see what types of models you will want to consider at the analysis stage.

Methods and Model Building

Results and Discussion

Appendix

```
knitr::opts_chunk$set(fig.height = 2, fig.width = 6,
                      echo = FALSE, fig.align = 'center',
                      message = FALSE, number_sections = FALSE)
library(alr4)
library(tidyverse)
library(ggplot2)
library(cowplot)
library(kableExtra)
library(tidymodels)
library(patchwork)
library(MASS)
library(GGally)
library(dplyr)
library(hms)
library(corrplot)
## pull in data and clean
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_avgTop3")
```

```r
## remove missing values
weather_missing = which(is.na(results$NOAA_ID) | is.na(results$rel_hum))

results_missing = which(results$world_record == "" |
                        results$standing_record == "" |
                        results$time_1st == "")
results = results[-c(weather_missing, 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)
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)
results$time_avgTop3_s = as.numeric(results$time_avgTop3)

## add proximity variable
results$proximity = ifelse(results$dist_from_loc <= 10, "close", "far")

## percent deviation from world and standing record
results = results %>%
  mutate(percent_world_1st = (time_1st_s - world_record_s) /
           world_record_s * 100,
         percent_standing_1st = (time_1st_s - standing_record_s) /
           standing_record_s * 100)

## make overall date of competition variable
results = results %>%
  mutate(date = make_date(year, month, day))
# create proximity variable
table(results$proximity)
#summary(results$percent_standing_1st)
ggplot(data = results, aes()) +
  geom_histogram(aes(x = percent_standing_1st)) +
  xlab("Percent Deviation between First Place and Standing Record") +
  ylab("Count")
weather = c("air_temp", "dew_point", "wind_speed",
            "adj_wind_speed", "rel_hum", "clouds_OKTA", "solar_rad",
            "heat_index", "simp_wet_bulb", "wet_bulb")
temp = results[results$proximity == "close" ,c(weather, "percent_standing_1st")]
M.under = cor(temp, use = "na.or.complete")
corrplot::corrplot(M.under, method = "ellipse", type = "lower",
```

```
         tl.cex = 0.6, cl.cex = 0.5, mar = c(0,0,0,0))
temp = results[results$proximity == "far" ,c(weather, "percent_standing_1st")]
M.over = cor(temp, use = "na.or.complete")
corrplot::corrplot(M.over, method = "ellipse", type = "lower",
         tl.cex = 0.6, cl.cex = 0.5, mar = c(0,0,0,0))
ggpairs(data = results[, c("air_temp", "dew_point", "wind_speed", "heat_index", "simp_wet_bu
  theme(strip.text.x = element_text(size = 5),
        strip.text.y = element_text(size = 4))
## histogram of the distances
ggplot(data = results, aes()) + geom_boxplot(aes(x = dist_from_loc))

## proximity vs. response
ggplot(data = results, aes()) + geom_boxplot(aes(x = proximity, y = percent_standing_1st))

## subset dataframe based on values in dist_from_loc
under10 <- subset(results, proximity == "close")
above10 <- subset(results, proximity == "far")

## weather predictor scatterplots
ggpairs(data = under10[, c(weather[1:5], "percent_standing_1st")])
ggpairs(data = above10[, c(weather[1:5], "percent_standing_1st")])

ggpairs(data = under10[, c(weather[5:10], "percent_standing_1st")])
ggpairs(data = above10[, c(weather[5:10], "percent_standing_1st")])

ggpairs(data = results[, c(weather[1:5], "percent_standing_1st")])
ggpairs(data = results[, c(weather[5:10], "percent_standing_1st")])
## count table of sex, competition, and distance variables
cat.counts = as.data.frame(table(results$distance, results$competition, results$sex))
cat.counts = cat.counts %>% pivot_wider(names_from = c("Var1", "Var3"), values_from = "Freq"
cat.counts$Var2 = c("Common Wealth", "Diamond League", "Gold Marathon",
                    "Olympic", "World Champ.", "World Cup")

table = cat.counts %>% kable(col.names = c("Comp.", "10K", "20K", "3K", "50K", "5K", "Mar.",
                                           "10K", "20K", "3K", "50K", "5K", "Mar."), booktab
  add_header_above(c(" ", "Men" = 6, "Women" = 6)) %>%
  kable_styling(latex_options = "striped", full_width = FALSE)
table
ggplot(data = results, aes()) +
  geom_boxplot(aes(x = competition, y = percent_standing_1st)) +
  xlab("Competition") +
  ylab("Percent Deviation") +
  theme(axis.text.x = element_text(size = 7))
ggplot(data = results, aes()) +
  geom_point(aes(x = date, y = percent_standing_1st, color = distance),
             alpha = 0.5) +
```

```
  facet_grid(rows = vars(sex)) +
  xlab("Date") +
  ylab("Percent Deviation") +
  labs(color = "Distance")
ggplot(data = results) +
  geom_point(aes(x = heat_index, y = percent_standing_1st, color = distance),
             alpha = 0.5) +
  xlab("Heat Index") +
  ylab("Percent Deviation") +
  labs(color = "Distance")
```