Final Project Proposal

Date: 11/10/25

Basics

For this week, the roles are as follows:

- Team Lead: Meghan Holden
- Recorder: Krisha Bugajski
- Spokesperson: Zachary D'Urso

The **preliminary project title** is "*Predicting Marathon Average Finishing Times: The Impact of Weather and Air Quality on Marathons*."

Background and Question

Research Question

Can weather and environmental factors—such as temperature, wind speed, precipitation, dew point, visibility, sea level pressure, and air quality—predict average marathon finishing times across different performance groups (elite, competitive, average, and slow runners) and genders?

Need / Niche

By understanding how environmental factors affect average finishing times across different performance groups, this research can help:

- prepare security details and medical logistics along the course based on assumptions of where runners will be
- event organizers to prepare runner waves, hydration and cooling stations in the most effective ways
- runners to prepare their pacing and set their expectations as a result of given weather conditions

This serves a unique, cross-disciplinary niche within sports analytics, environmental science, and public health.

Why It’s Worth Exploring

The Boston Marathon, New York City Marathon, Chicago Marathon, and Berlin Marathon, are some of the largest marathon races in the world, part of the World Marathon Majors. If we can understand how weather and air quality affect performance for different groups of runners, we can help race organizers and directors improve race day plans, safety planning, and athlete preparation. In addition, the insights from this study can contribute to research on how climate and pollution trends affect endurance performance and outcomes in long-distance running events.

Novelty

There has been previous research on the effects of temperature on marathon finish times, as well as some investigations into pollutants. Ely, Cheuvront, Roberts and Montain (2007) found that for both men and women, regardless of marathon performance/skill group, between 5 and 25℃, there is a decrease in overall marathon performance with slower runners more greatly affected by the temperature increase. Marr and Ely (2010) found marathon performance for women, specifically, was impacted by an increase in PM_{10}.

However, there has not been much research that has modeled the impact of multiple environmental predictors, including weather, environmental conditions and air quality altogether. There have been few studies that have analyzed the effects across different performance groups. So the question is not new, but the specific outcome with the predictors, air quality data, and subgroups is original and meaningful.

Hypothesis

H_0 : Weather, and air quality have no significant relationship with the average finishing times across marathon performance groups or genders.

H_1 : Poor weather conditions (e.g., higher temperatures, dew point wind speed, precipitation, sea level pressure, or worse air quality) are associated with slower average finishing times, particularly for non-elite runners.

To isolate environmental effects, we will include a control variable for the introduction of supershoes, which helps account for performance improvements due to running shoe technology.

Predictions

- **Elite runners** will be least affected by environmental stressors due to higher fitness and experience (Ely, Martin, Cheuvront, & Montain, 2007).
- **Competitive runners** will show greater increases in average finishing times under harsher weather and environmental conditions (Vihma, 2010).
- **Slow runners** will experience the largest slowdowns under harsher weather and environmental conditions (Helou et al., 2012).
- **Female runners** will show slightly smaller performance declines under adverse weather and environmental conditions compared to males (Vihma, 2010).
- **Moderate temperatures, low humidity, and clean air** will produce the fastest finishing times across all categories (Gasparetto & Nesseler, 2020).

Data & Analysis

The following datasets provide the relevant data for the predictor and response variables covering the relevant geographic regions and temporal span as the marathon data. The focus of this analysis will be from 1996 through 2025, data that is available in the datasets earlier than 1996 will not be included.

Datasets:

1. **Marathon Results Data**

   Source: Boston Marathon Data Project – GitHub

   - Contains race results by year from 1970 through 2019.

   Source: New York City Marathon – Kaggle

   - Contains race results by year from 1970 to present.

   Source: Chicago Marathon – Kaggle

   - Contains race results by year from 1996 to 2023.

   Source: Berlin Marathon – Kaggle

   - Contains race results by year from 1974 to 2019.

2. **Weather Data**

   Source: Weather Underground

   - Provides average daily min, max, and average temperature, precipitation, dewpoint, wind, visibility and sea level pressure in Chicago, Boston and New York City.

Source: Meteostat

- Provides average daily temperature, precipitation, wind, and sea level pressure in Berlin.

Source: Extreme Weather Watch

- Provides min and max temperature for Berlin

Source: Weatherspark.com

- Provides a full day's collection of dew point and visibility data for Berlin, used to manually calculate average daily dew point and visibility.

3. **Air Quality Data**

Source: EPA Daily Air Quality Tracker

- Provides daily AQI (Air Quality Index) values and primary pollutant information.

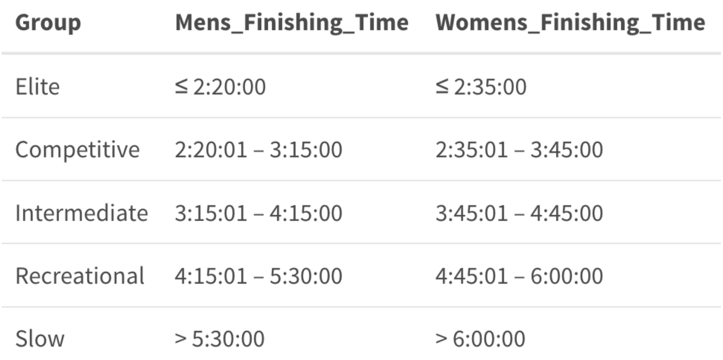
Source: EPA Daily Air Quality Data Download Portal.

- Provides daily AQI (Air Quality Index) values and specific pollutant measurements including NO_2, O_3, $PM_{2.5}$, PM_{10}, and CO.
- These pollutants are part of the EPA's official race-safety guidelines for event directors and will allow for a deeper analysis of how specific air quality components relate to finisher rates.

Response variable

The response variable will be the average finishing time (format showing hours, minutes, and seconds). This variable will reflect the overall race performance in a given year.

Runners will first be divided into male and female groups. Nonbinary runners will be grouped with female runners so that their data is not underrepresented in the model. The decision to group nonbinary runners with female runners is related to their identical Boston Marathon Qualifying Standards (BAA, 2025). Runners will then be further grouped into performance groups as shown below (PR Run & Walk, 2025):



Group	Mens_Finishing_Time	Womens_Finishing_Time
Elite	≤ 2:20:00	≤ 2:35:00
Competitive	2:20:01 – 3:15:00	2:35:01 – 3:45:00
Intermediate	3:15:01 – 4:15:00	3:45:01 – 4:45:00
Recreational	4:15:01 – 5:30:00	4:45:01 – 6:00:00
Slow	> 5:30:00	> 6:00:00

We adapted the PR Run & Walk marathon time ranges to remove small overlaps and gaps between groups. This keeps the categories (Elite, Competitive, Intermediate, Recreational, Slow) consistent with the source but makes them statistically cleaner and easier to use for modeling. Minor adjustments ensure the groups are mutually exclusive and continuous while maintaining the intent of the original data.

This approach will give us multiple observations per year by marathon (marathon × year × gender × performance group), increasing the dataset size for modeling.

Predictor variables

Type	Variable
Weather Variables	Temperature (°F), Precipitation (in), Average Dew Point (°F), Wind (mph), Visibility (statute miles), Sea Level Pressure (inHg)
Air Quality Variables	AQI Value, Main pollutant (categorical: PM2.5, PM10, O_3, NO_2, Pb, CO, SO_2)
Temporal Variables	Year, day-of-week (race day)

Analysis plan:

Data Collection & Cleaning:

- We will combine the race results with weather and air quality data for each race year, using the conditions from race day.
- We will ensure all units are consistent between regions and races.
- We will group runners per year to performance groups using pre-determined cutoffs.
- In instances where the original data appears incorrect, additional searches for data will be investigated to fill in, otherwise we will fill in missing numbers with median values for numeric features and “Unknown” for categories to keep the dataset clean and consistent.

Feature Engineering:

- We will compute the average finishing times per group per year.
- We will create interaction terms like *temperature* × *dew point* and *AQI* × *temperature.*
- We will normalize finishing times by year and city to account for different courses.
- We will normalize the weather variables across the races.

- We will create a control variable to account for the introduction of “super-shoes” for marathon results starting in 2018 (the year following their official public release). We will code 0 before 2018 and 1 for 2018 onward.
    - Research from Cornell (Guinness et al, 2020) looked into performance of elite and subelite athletes from 2015 to 2019 and found athletes wearing Nike Vaporflys had improved performance of up to 3.5 minutes for women, and up to 3.9 minutes for men. Paradisis, Zacharogiannis, Bissas, & Hanley (2023) confirmed improved results for slower runners as well; therefore, we will use this control variable to account for their introduction.

Modeling:

- We will start with a Multiple Linear Regression model to see basic relationships, and then compare it with a Gradient Boosting model, such as XGBoost or LightGBM, to see more complex patterns.
- The dataset will be split into training and test sets (e.g., 80/20) to evaluate model performance, using metrics such as R^2, MAE, and RMSE using cross-validation.

Evaluation & Interpretation:

- We will then visualize actual vs. predicted average finisher times, review feature importance, and check residuals for bias.
- We will finally interpret how factors like temperature, wind, humidity, and air quality influence average finish times among each of the performance groups.

Potential pitfalls:

The first marathon race with the modern distance of 26.2 miles was run at the 1908 London Olympics. Since then, shoe technology and training plans have come a long way, resulting in faster and faster marathon times. Because of this, our data will be limited to 1996 through 2025, inclusive, and we will introduce a control variable for super shoes used in marathons beginning in 2018. Another pitfall we will experience is the impact external factors have on the races, reducing the data available. These include events like the COVID-19 pandemic, Hurricane Sandy hitting New York City in 2012, and the tragic terrorist attack at the finish line of the 2013 Boston Marathon. Because of the potential pitfall of a small sample size as a result of only focusing on approximately 30 years, and to reduce the impact of external factors leading to limited or no race results, we expanded the number of marathons being included in this model from one to four. There is still the possibility that aspects of the marathons have changed over time that could impact results (ex: faster cutoff times needed to participate in the race).

We may experience difficulty accessing historic air quality data for Berlin, in this event we will search for alternative data sources, and if needed we will attempt to replace the Berlin marathon with another US based marathon.

Multicollinearity could be an issue, as some environmental factors may be highly correlated with one another. For example, we may see a correlation between dew point and precipitation, or precipitation and air quality. In order to assess this, we will create a correlation matrix, from which we may determine whether to drop a feature. Also, if any data is missing related to a race or environmental conditions for the day, it could throw off the results, therefore we will first try to find the data using supplemental sources, then apply imputation where appropriate.

How will you know if your question is answered?

We will know if our question is answered if our created predictive model is capable of explaining a majority of the variation in the marathon average finishing times. Additionally, the model should help us identify the most significant predictors.

How will you know if your hypothesis is supported?

We will know if our hypothesis is supported if our environmental factors display statistically significant impacts on the marathon average finishing times.

Technical Details

Coding Language:

We plan to code for our project using R for predictive modeling and visualization purposes.

Other resources:

We will utilize several downloaded datasets, RStudio for coding, a shared GitHub repository to manage our code, a shared google drive, and academic sources for research purposes.

Group's GitHub repo link:

github.com/BugajskiSharp/capstone-project-team-c

References

- Average marathon times by age & ability. =PR= Run & Walk. (2025, October 3). https://prrunandwalk.com/blogs/news/average-marathon-times#:~:text=Elite%20Runners,size%20of%20hot%20air%20balloons
- Boston Athletic Association. (n.d.). *Qualifying for the Boston Marathon. Boston Athletic Association.* https://www.baa.org/races/boston-marathon/qualify/
- Ely, M. R., Martin, D. E., Cheuvront, S. N., & Montain, S. J. (2007). *Effect of ambient temperature on marathon pacing is dependent on runner ability. Medicine & Science in Sports & Exercise, 39*(2), 221–229.
- Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). *Impact of weather on marathon-running performance. Medicine & Science in Sports & Exercise*, 39(3), 487–493. https://doi.org/10.1249/mss.0b013e31802d3aba
- Gasparetto, T., & Nesseler, C. (2020). *Hot and polluted, but comfortable: The effects of weather and pollution on running performance in marathons. Journal of Sports Economics*, 21(10), 1038–1062.
- Guinness, J., Bhattacharya, D., Chen, J., Chen, M., & Loh, A. (2020). *An observational study of the effect of nike vaporfly shoes on marathon performance. arXiv preprint* arXiv:2002.06105.
- Helou, N. E., et al. (2012). *Impact of weather on marathon-running performance. PLoS ONE, 7*(5), e37407.
- Marr LC, Ely MR. *Effect of air pollution on marathon running performance. Medicine & Science in Sports & Exercise*. 2010 Mar;42(3):585-91. doi: 10.1249/MSS.0b013e3181b84a85. PMID: 19952812.

- Paradisis, G. P., Zacharogiannis, E., Bissas, A., & Hanley, B. (2023). *Recreational runners gain physiological and biomechanical benefits from super shoes at marathon paces*. *International Journal of Sports Physiology and Performance, 18*(12), 1420-1426. https://doi.org/10.1123/ijspp.2023-0115
- Vihma, T. (2010). *Effects of weather on the performance of marathon runners. International Journal of Biometeorology, 54*(3), 297–306.