Preliminary Project Proposal

Date: 11/03/25

Basics

For this week, the roles are as follows:

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

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

Background and Question

Research Question

Can weather and environmental factors—such as temperature, humidity, wind speed, precipitation, and air quality—predict average finishing times across different performance groups (elite, competitive, average, and slow runners) in the Boston Marathon?

Need / Niche

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

- prepare medical officials and medical tents with support needs
- event organizers to prepare 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 is one of the largest races in the world. If we can understand how weather and air quality affect performance for different groups of runners, we can 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. However, there are not many that have modeled the impact of multiple environmental

predictors, especially including air quality. Also, there have been few studies that have analysed 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 factors have no significant relationship with the average finishing times across runner performance groups in the Boston Marathon.

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

Prediction

- Elite runners will be relatively resilient to environmental stressors, showing smaller performance impacts.
- Competitive, average, and slow runners will show progressively larger increases in average finishing times under adverse conditions.
- Moderate temperatures, low humidity, and clean air will correspond to faster average finishing times across all groups.

Data & Analysis

Datasets:

1. **Boston Marathon Results Data (through 2019)**

   Source: Boston Marathon Data Project – GitHub

   - Contains race results by year, through 2019, and allows creation of performance groups using percentile-based cutoffs.

   Source: Boston Marathon Results

   - Contains race results by year, from 2001 to present, and allows creation of performance groups using percentile-based cutoffs.

2. **Weather Data**

   Source: NOAA / Climate.gov Historical Weather Data

   - Provides average daily temperature, humidity, wind speed, and precipitation for the Boston area.

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.

   - We also found a more detailed EPA dataset that includes specific pollutant measurements — 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.

These are good datasets because they cover the same geographic region (Boston) and they have the same temporal span as the marathon data (through 2019). Data from 2013 will be omitted due to the tragic terrorist attack impacting race completion.

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 be grouped into:

Elite: Top 5%

Competitive: 5-20%

Average: 20-80%

Slow: 80-100%

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

Predictor variables

Type	Variables
Weather Variables	Average temperature (°C), humidity (%), wind speed (m/s), precipitation (mm)
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)

Tentative 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 group runners to performance groups using percentile cutoffs per year.
- 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* × *humidity* and *AQI* × *temperature.*

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.
- Model performance will then be evaluated with 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:

Marathon only occurs once a year, so there is potential for a small sample size to be a pitfall. Multicollinearity could be an issue as some environmental factors may be highly correlated with one another. An additional pitfall could be aspects of the marathon that have changed over time (ex: faster cutoff times needed to participate in the race). If any data is missing related to the race or environment, it could throw off the gathered results, as we would need to find a resolution to combat this (ex: imputation). Another pitfall we will experience is that in 2013, there was a tragic terrorist attack on the route of the Boston Marathon, which claimed lives and injured many. As a result, some runners were not able to finish the race, potentially impacting the strength of our overall results, specifically in the 2013 marathon. For this reason, we plan to omit the 2013 data in our project.

Additionally, because of the COVID-19 pandemic, the 2020 race was cancelled, and the 2021 race was postponed to October instead of occurring on Patriots' Day (third Monday of April). This may skew the results data because of differences between summer and winter training, and the potential impact that an increased rate of respiratory illness may have on training. The 2021 event was also limited to 20,000 runners (10,000 fewer than normal recent years), which may impact the subgroup results.

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 at least three downloaded datasets, RStudio for coding, a shared GitHub repository to manage our code, and academic sources for research purposes.

Group's GitHub repo link:

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