Predicting Marathon Average Finishing Times:

The Impact of Weather and Air Quality on Marathons

Final Written Deliverables

By

Krisha Bugajski-Sharp, Zachary D'Urso, Meghan Holden

* Each author contributed equally to the design, coding & development, analysis, and writing of this project

December 17, 2025

Table of Contents

Abstract

Running a marathon is hard work on its own, but the challenge becomes even harder when environmental factors such as weather and air quality come into play. Some years have perfect conditions where the weather is cool, while other years are hot, rainy, and have heavy pollution. Running has also had many changes in the past decade, including the introduction of "supershoes," making it harder to tell whether faster or slower finishing times are due to environmental factors or the shifts in technology. This raises the question: How much do weather and air quality affect marathon performance, and can we predict finishing times before the race even begins?

In this project, we analyze thousands of runners from Boston, Chicago, and New York City marathons by combining official race results with historical Weather and EPA air quality data. After cleaning, preprocessing, and creating meaningful features, we compare several types of modeling to understand which conditions matter the most. We start with a linear model to set baseline relationships, then use feature selection tools such as decision trees and LASSO, and finally build two XGBoost models, one with newly engineered features and the other without, so it can learn complex patterns automatically. Our goal is simple: identify which environmental conditions slow runners down, who is affected the most, and which modeling strategies give us the most accurate and interpretable predictions.

Our findings show that environmental conditions do matter, especially for slower and recreational runners, while elite runners remain the least affected. Weather and air quality

variables improve prediction accuracy a lot when used with strong modeling techniques. Our strongest model, XGBoost with engineered features, outperforms the linear model and XGBoost model with raw features, and highlights which conditions slow runners down the most.

Background and Question

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 that 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. Overall, there have been few studies that have analyzed the effects across different performance groups. Also, marathon results have been influenced by non-environmental factors such as shoe technology. These trends can sometimes mask the roles of environmental factors, and that is why we were motivated to make these model-based approaches that are capable of isolating environmental effects.

Research Question

This research looks at how environmental factors like weather and air quality affect marathon finishing times for different groups of runners from 1996 to 2024 to answer the 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?*

Results of this project can provide helpful insights that can guide organizers to plan security, medical support, and hydration stations, while allowing runners to pace themselves and adjust their expectations. Looking at multiple environmental factors together can improve race day planning, enhance overall safety, and deepen our understanding of how climate and pollution influence performance in major marathons like Boston, New York, and Chicago.

Hypothesis

We hypothesize that poor weather conditions, such as 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.

Predictions

Based on prior research, we expect:

- **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

Data Acquisition

The following datasets provide the relevant data for the predictor and response variables covering the relevant geographic regions and temporal span of the marathon data. The focus of this analysis will be from 1996 through 2024, and data that is available in the datasets earlier than 1996 will not be included.

Although Berlin data was initially included, it was later separated and used as a case study due to substantial missingness, rather than being incorporated into the main modeling dataset. A data dictionary describing the features is available in Appendix 1.

Datasets:

1. **Marathon Results Data**

Marathon data was pulled directly from Kaggle and GitHub in csv form. The marathon datasets contain individual race results, including key identifiers such as *year* and *marathon*, along with the grouping variable *gender.* It also includes *chip_time,* which we use to create our outcome variable, *avg_chip_time*, as well as performance subgroups. Having these variables within the dataset also reduces the need for additional dataset joins.

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 2024.

Source: Chicago Marathon – Kaggle

- Contains race results by year from 1996 to 2023.

Source: Berlin Marathon – Kaggle

- This will act as a case study. Contains race results by year from 1974 to 2019.

2. **Weather Data**

The Weather dataset was manually compiled and includes 115 observations and 11 variables. It is important to note that the weather measurements may not perfectly represent on-course conditions, as monitoring stations were sometimes located away from race routes. This dataset contains the exact marathon *Date* for each year, which allows us to find exact weather data for that given day. It also contains key identifiers, *Year*, and *Marathon,* which allowed it to be merged with the marathon dataset. The dataset also contains the predictor variables *High Temp, Low Temp, Day Average Temp, Precipitation, Average Dew Point, Max Wind Speed, Visibility, and Sea Level Pressure.*

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 temperatures 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**

The Air Quality dataset was also manually compiled and includes 115 observations and 10 variables. It is also important to note that the air quality measurements may not perfectly represent on-course conditions, as monitoring stations were also often located away from race routes. This dataset contains the exact marathon *Date* for each year, which allows us to find exact air quality data for that given day. It also contains the key identifiers, *Year,* and *Marathon*, which allowed it to be merged with the marathon dataset. The dataset also contains the predictor variables *Overall AQI Value, Main Pollutant, Ozone, PM10, PM2.5, NO2,* and *CO.*

Source: EPA Daily Air Quality Tracker

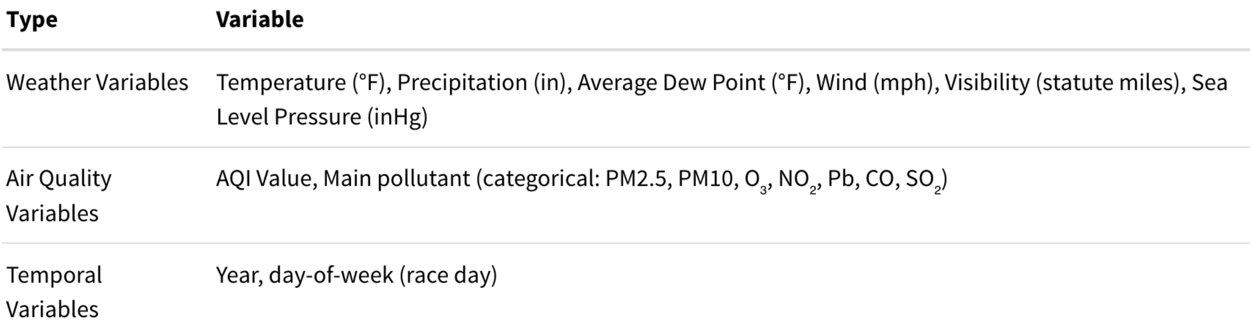
- Provides daily AQI (Air Quality Index) values and primary pollutant information.

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.

Original Predictor variables

Table 1. Contains all original predictor variables along with their units, prior to preprocessing and feature engineering.



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)

Response variable

The response variable is *average_chip_time*, which is calculated by converting individual chip times to seconds and then averaging within each performance group by marathon, year, and gender. This variable represents overall race performance for a given group in a given year.

Cleaning

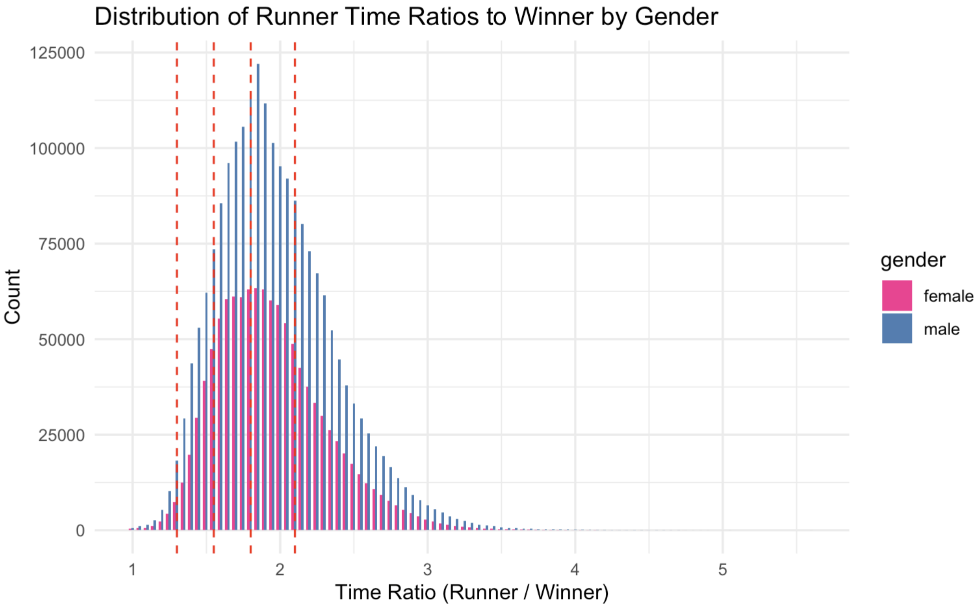
The marathon datasets from Boston, New York City, and Chicago were first standardized to ensure they were consistent across all years and marathons. Then variable names were cleaned, and only the necessary columns were chosen, including *year, marathon, gender,* and *chip_time.* Weather and air quality datasets were also standardized in the same way, with variables renamed for consistency.

All marathon datasets were then combined using row-wise binding. Only the years from 1996 to 2024 were retained, making sure the analysis period was consistent. We addressed missing or invalid *chip_time* values by using a custom function called (`clean_chip_time`) that converted time to a uniform HH:MM:SS format and then replaced entries that were not valid with NAs. These clean chip times were then converted to total seconds (`chip_seconds`) for easier calculations. There were only three rows that had missing chip times, which were removed safely because we used group averages in the analysis.

Next, gender labels were standardized to male, female, or unknown. Nonbinary runners were included under female due to their small number, which is related to their identical Boston Marathon Qualifying Standards (BAA, 2025). The small number of unknown genders, consisting of 56 rows, was removed to reduce noise, leaving only male and female runners.

Performance-based subgroups were created using the *time_ratio,* which is each runner's finishing time divided by the winner's time for the same year, marathon, and gender. Subgroup thresholds were found by the histogram of time ratios, showing natural clustering of runners into five categories, including elite, competitive, average, recreational, and slow. Figure 1 shows how these thresholds were chosen based on the distribution of data rather than percentiles. Subgroups were further divided by gender so we could maintain meaningful comparisons.

Figure 1. Histogram showing the distribution of runner time ratio values (finishing time/winner time), shown separately for males and females. Vertical dashed red lines mark the chosen subgroup cutoffs (1.30, 1.55, 1.80, and 2.10).



Average finishing times were computed for each subgroup, year, marathon, and gender, resulting in a total of 770 observations with six core variables. There were some years and genders missing from specific marathons, such as NYC male runners in 2024. This however does not impact the analysis focus, which is on maximizing the coverage of finishing times.

Finally, the cleaned marathon data was merged with the standardized weather and air quality datasets using *year* and *marathon* as key identifiers. We added a column `n` to show the number of runners found in each group, allowing us to enable potential weighting for future analyses.

Underlying Assumptions:

1. Missing chip times are random, and removing a few rows does not cause any bias within the subgroup averages.
2. Weather and air quality measurements from nearby stations are able to reasonably show approximate course conditions.
3. Time ratio-based performance categories are able to reflect meaningful distinctions in a runner's ability.
4. Grouping nonbinary runners with female runners should not cause a large variation in the female finishing times, as their qualifying times for races, like the Boston Marathon, are the same as female runners.

Data Exploration

Before moving on to modeling, it is important to understand the structure, distribution, and all the relationships within the final dataset. Data exploration allows us to find patterns, detect any anomalies in our dataset, and see how the predictor variables relate to the outcome variable. The following visualizations and summary tables are able to provide insights into performance trends across marathons, years, and gender subgroups, as well as how environmental factors may influence finishing times.

Figure 2 shows an initial glimpse at how marathon finishing times change alongside average temperature, broken down by marathon. Each marathon has a positive relationship between the increase in average temperature and an increase in the average finishing time.

Figure 2. Relationship between average temperature and finishing time by each marathon.

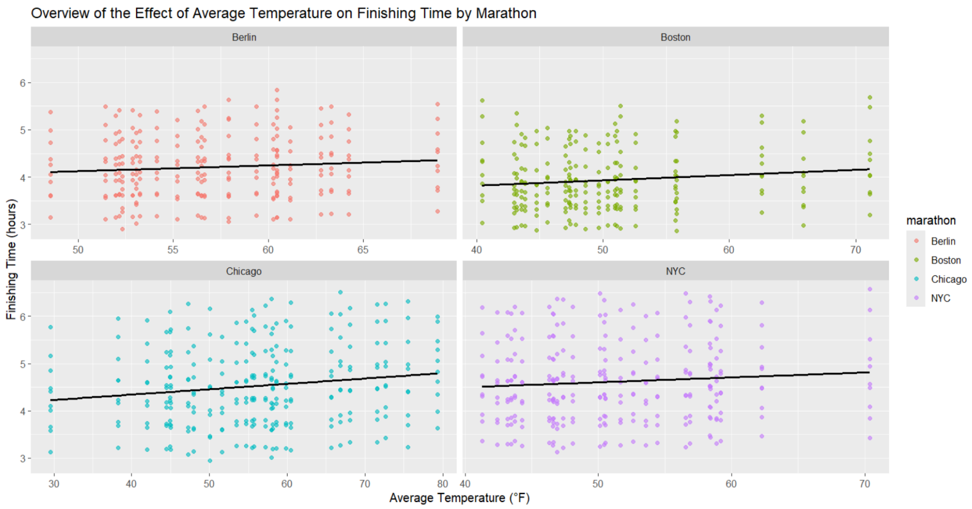
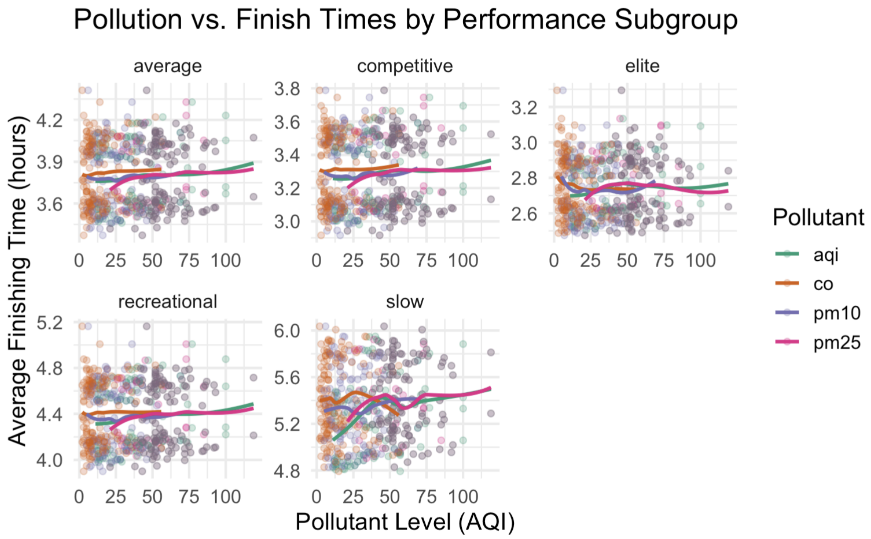


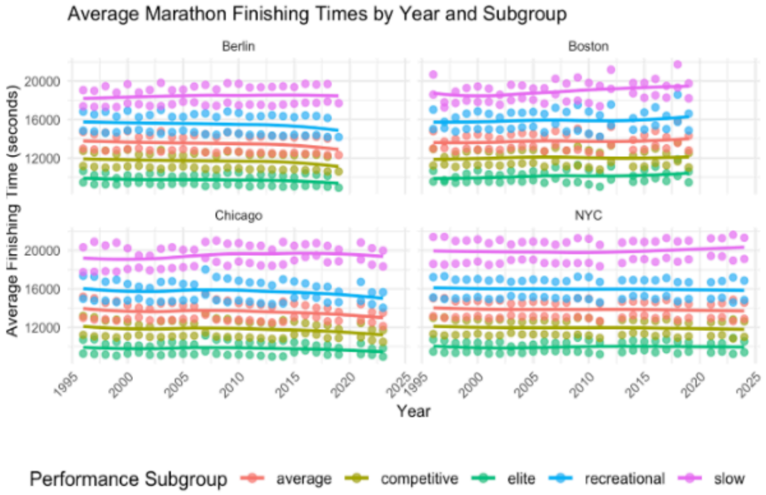
Figure 3 shows initial exploration into how marathon finishing times change with different pollution levels from across performance groups using the original data. As pollution, especially PM2.5, increases, finish times generally get slower. Elite runners are the least affected, while slower runners show the biggest delays.

Figure 3. Scatterplot showing the relationship between pollutant levels and finishing times by subgroups.



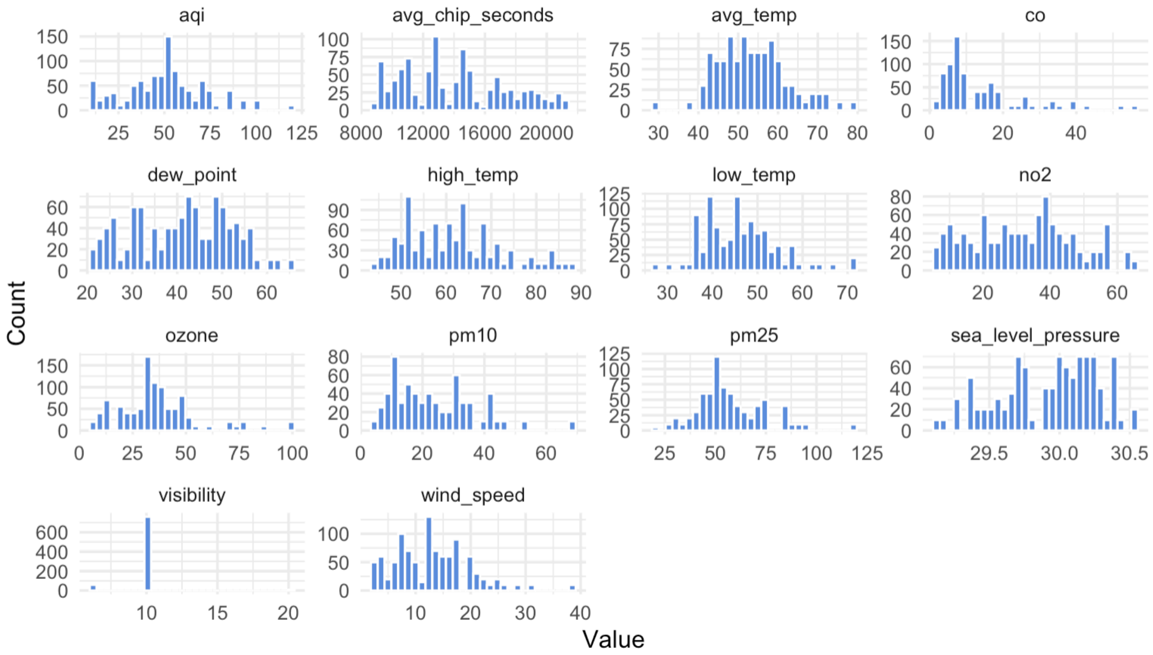
Performance-based subgroups were created using the *time_ratio,* which is each runner's finishing time divided by the winner's time for the same year, marathon, and gender. Figure 4 illustrates how these subgroups showed an apparent increase in average finishing times over the years, specifically in Chicago and Berlin (which, as mentioned above, was used in a separate case study of the final model).

Figure 4. Scatter plots by marathon and subgroup. Colors represent the different performance groups.



We further looked at histograms for all the continuous variables to see distribution shapes and confirm that finishing time conversions were correct, as seen in Figure 5. Several environmental variables did display expected right-skew (e.g. PM2.5, wind speed), which was taken into consideration for the linear model.

Figure 5. Histograms of Continuous Variables



After making sure all variable types were correct, so that numerical features were stored as numeric and categorical, binary, and logic features were converted to factors, we ran a correlation matrix to check for multicollinearity amongst all numeric features, including the ones just engineered, as seen in Figure 6. We found that the original features, *high_temp* and *low_temp,* were highly correlated with *avg_temp,* and *aqi* was highly correlated with *pm2.5,* as seen in Table 2.

We also created several interaction terms to try to capture potential combined effects of environmental and demographic factors on marathon performance, as summarized in Appendix 2. Recognizing that we might drop them due to high correlations, we generated a correlation matrix to assess. For instance, we created interactions between average temperature, dew point, wind speed, precipitation, and air quality to reflect weather and pollution influences that might not be fully captured by individual variables alone; however, upon inspection of the correlation matrix in Figure 6, we decided to remove the highly correlated interaction terms. For example, high temperature and low temperature were both highly correlated with average temperature, so we dropped the high and low temperatures and kept the average temperature. The list of original features and interaction terms that were removed from the final dataset can be found in Table 2.

Figure 6. Correlation matrix that includes the newly engineered features.

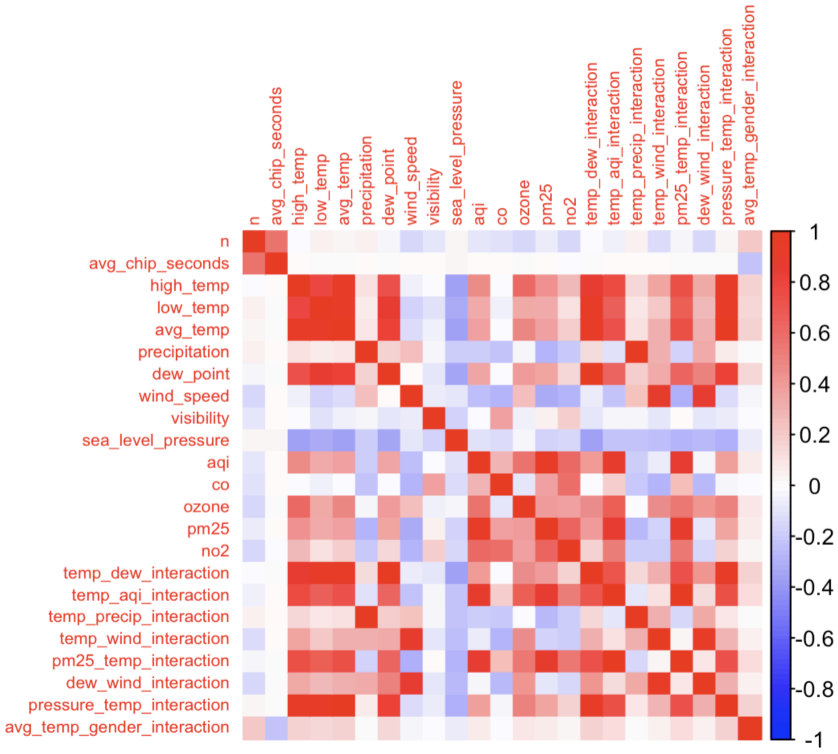


Table 2. Features to drop

Features to Drop	**Reason or (High Correlation with)**	**Correlation**
high_temp	avg_temp	0.93
low_temp	avg_temp	0.92
aqi	pm25	0.94
main_pollutant	Categorical version of aqi	**NA**
temp_dew_interaction	avg_temp **&** dew_point	0.94 **&** 0.96
temp_precip_interaction	precipitation	0.99
temp_wind_interaction	wind_speed	0.89
pm25_temp_interaction	pm25 **&** temp_aqi_interaction	0.89 **&** 0.97
dew_wind_interaction	wind_speed	0.84
pressure_temp_interaction	avg_temp	1.00

Models

Pre-processing and dimensionality reduction / feature engineering

A series of preprocessing, feature engineering, and reduction, and supervised modeling approaches were utilized to develop a deeper understanding of how weather and air quality both influence marathon finishing times across the Boston, Chicago, and NYC marathons. The final dataset contained numerous environmental variables, categorical descriptions for the marathons, and various feature-engineered interaction terms.

We started by converting all categorical variables to factors, including *marathon*, *gender*, *subgroup*, and *supershoe,* since the linear model and XGBoost models require the encoding of categorical variables. Our numeric and continuous predictors consisted of temperature, wind speed, dew point, CO, PM2.5, and ozone.

We assessed missingness and removed PM10 because it had 320 missing values and would introduce noise rather than signal, while PM2.5, which is an important pollutant with only 70 missing values, was retained and imputed using KNN (k = 5) to preserve structure and variability without getting rid of any rows.

Features were scaled for the linear model due to skew and non-normal distribution (Appendix 3 and Appendix 4), but no transformations were required for the XGBoost models, as tree-based methods are more robust at dealing with non-normal distributions in the data.

There were several newly engineered features created in order to capture the most valuable interactions (Appendix 4). As mentioned above, after investigating our new features in a correlation matrix (Figure 6), we decided to select which highly correlated variables to remove. After all features were removed, as described in Table 2, the remaining engineered features were the supershoe control variable, temperature AQI interaction (temperature x AQI), and average temperature and gender interaction (temperature x gender).

No unsupervised feature engineering was conducted for this model. PCA was considered as an unsupervised approach for additional feature engineering, but ultimately discarded as our environmental predictors carry clear meanings. PCA would combine our environmental features to create new components that would be challenging to interpret in terms of overall marathon performance. Maintaining interpretability is important because we want to understand the effects of environmental conditions on finishing times, and not just maximize predictive accuracy.

A decision tree and LASSO regression were used as supervised methods for feature engineering to determine if the model would benefit from converting any continuous feature into a binned feature. The decision tree suggested binning windspeed, average temperature, PM2.5, ozone, and the temperature/AQI interaction term.

We tested these in linear regression models to identify which binned features are best to carry forward. Results from the models showed the model with ozone and PM2.5 binned has the lowest RMSE and MAE, suggesting higher accuracy, and one of the highest R^2 to explain variability. The model with just ozone binned has the highest R^2, but also slightly higher RMSE and MAE.

We verified these results using LASSO regression to determine whether the scaled version or the binned versions should be kept for modeling. The LASSO model retained ozone_bin, pm25_bin, and scaled_pm25 features. It showed that scaled_ozone shrank nearly to 0, meaning it did not contribute to predicting the outcome variable once binned features were included. Based on these results, and for a simpler and more interpretable model, we decided to retain the binned features for pm2.5 and ozone (Table 3) and remove the parent features for the next stage in the modeling process.

Table 3. Binned Features to Investigate

Feature Name	Binning Description
ozone_bin	High: ≥ 38 Low: < 38
pm25_bin	High: ≥ 54 Low: < 54

The final step before initiating our baseline model was to split the data into a training and test set. Due to the small size of the data set, we utilized a 90/10 training/test split to allow the model to have as much data to train on as possible.

Algorithm Selection

Our modeling plan was to first use a multiple linear regression model as a baseline, and then pivot to utilizing more nonlinear and flexible XGBoost models. The benchmark linear regression model allowed for the interpretation of the importance of all our features. Diagnostic plots in Appendix 5 show that model assumptions were met based on plotted residuals and investigation into leverage and variance. Test RMSE, MAE, and R-squared values for the linear regression model, shown in Table 5, were used as our baseline values to improve upon by utilizing several XGBoost tree-based gradient boosting models.

XGBoost models are able to automatically capture nonlinear relationships and interaction effects. The XGBoost model makes no strict distributional assumptions but requires numeric inputs (categorical variables must be encoded), is relatively robust to multicollinearity, and benefits from sufficient data to avoid overfitting through tuning.

This makes XGBoost a particularly fitting choice for our analysis, given that many environmental factors interact in complex ways. Additionally, XGBoost supports cross-validation for hyperparameter tuning and provides measures of variable importance, as well as compatibility with SHAP analysis for interpretable results. Using XGBoost allows us to assess model performance both with and without feature engineering.

Two baseline XGBoost models were built: one with feature engineering and one without, as seen in Table 4. Both the baseline models, which were trained without any hyperparameter tuning, showed substantial overfitting.

Table 4. Comparison of Training RMSE and CV RMSE

Model	Train RMSE	CV RMSE
XGBoost Baseline (feature engineered)	31.2820	145.6836
XGBoost Baseline (raw features)	30.6632	177.8890

We then created better-tuned XGBoost models by using hyperparameter-tuning loops to select the optimal parameters. We used these instead of guessing which settings would give us the best performance. These loops are able to automatically test several combinations using 5-fold cross-validation. They can then check how well the models predict on unseen data and record the best RMSE and number of boosting rounds. Then it can select the combination with the lowest cross-validated error. With the tuning loops, we get well-balanced models that are flexible enough to capture real patterns while being regularized enough to avoid memorizing all the noise.

After running both the XGBoost models, we found that the feature-engineered XGBoost model produced the best results. For this model, the best parameters were:

- max_depth = 4
- min_child_weight =5
- gamma = 0.1
- subsample = 0.8
- colsample_bytree = 0.8
- Eta = 0.1
- L1 and L2 penalties where alpha = 0 and lambda = 1
- Optimal number of boosting iterations = 1248

Final Model

After testing various models as previously discussed, we determined that our best model was the feature-engineered XGBoost model. While linear regression captured important baseline relationships, our goal was to further reduce prediction error. Among the models we tested, the feature-engineered XGBoost model performed best, as it more effectively captured interactions between performance groups and environmental conditions.

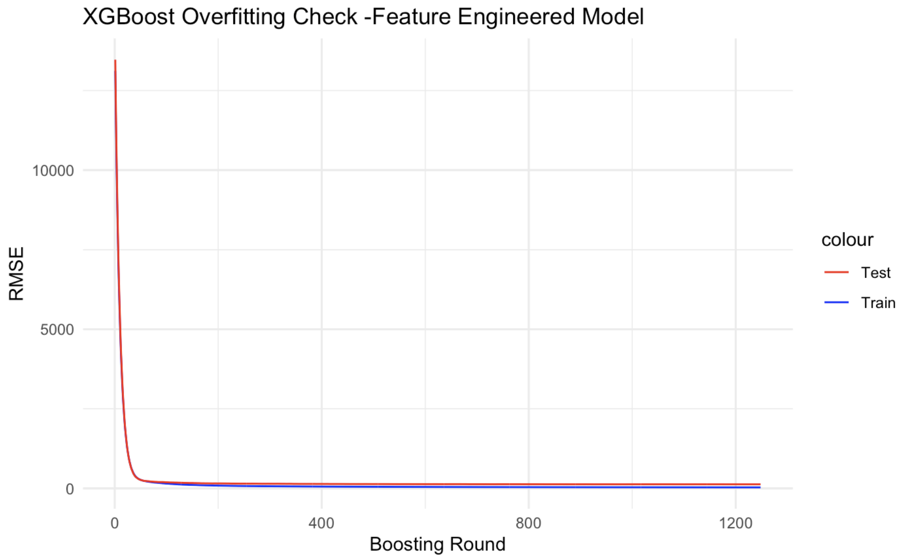
Table 5 shows the comparison of our models with XGBoost (feature engineered), championing our selected model due to the low RMSE and MAE.

Table 5. Model Prediction Comparisons on Test Data

Model	**Test RMSE**	**Test MAE**	**Test R^2**
Linear Regression (Baseline model with engineered features)	≈ 358.0687	274.0892	0.9892
XGBoost (feature engineered)	**≈ 127.4087**	**82.1464**	**0.9987**
XGBoost (raw features)	≈ 120.8608	90.1244	0.9987

Figure 7 shows the learning curve produced by plotting the training and test RMSE across the boosting rounds. The training RMSE is slightly below the test RMSE, suggesting mild overfitting; however, the gap is fairly small and fairly stable. This is consistent with the comparison of final RMSE values, where training RMSE is 31.3 and test RMSE is 127.4.

Figure 7. XGBoost Overfitting Check on the Feature-Engineered Model



We further tested our selected model on the Berlin data; the results indicate that the XGBoost engineered-feature model predicts average chip times for Berlin marathon runners with high accuracy on the subset of data with complete predictor information (Table 6). The R^2 of 0.9322 shows that over 93% of the variance in actual chip times is explained by the model. While the RMSE is approximately 817 seconds (or 13.6 minutes), which is six or seven times that of the original model's performance on the test data, this error is still practically useful given the amount of missing Berlin data.

Table 6. XGBoost Engineered-Feature Model Performance on Berlin Test Data

Metric	**Value**
RMSE	816.8284
MAE	616.8228
R^2	0.9322

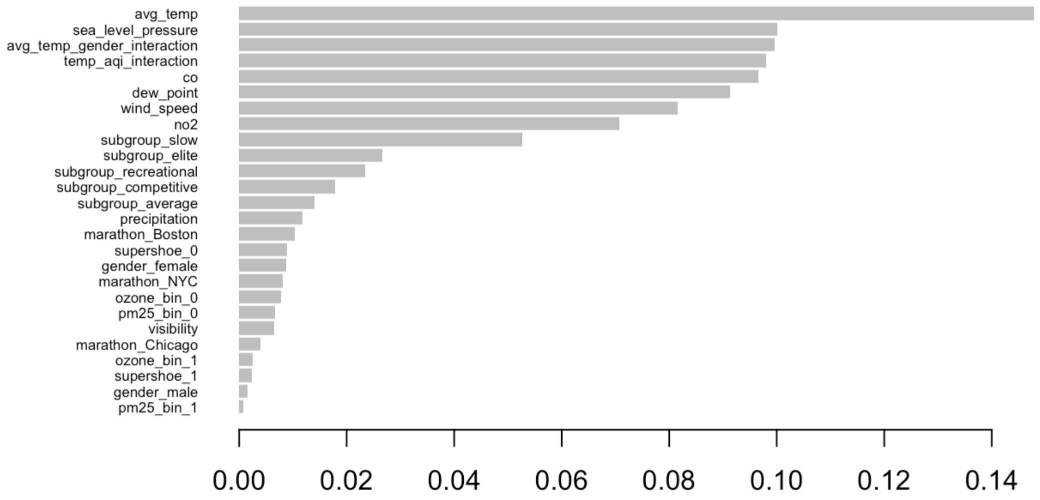
Out of 235 total Berlin observations, only 60 rows (or around 26%) were actually used for prediction. 175 rows (or around 74%) were dropped due to missing values in predictors. This is a substantial portion of the dataset, which means the evaluation metrics reflect model performance only on the complete-case subset.

Conclusions

Based on the comparison of test RMSE, MAE and R^2 (Table 5) across our models, as well as investigation into overfitting (Figure 7), the XGBoost model trained on the engineered features provides the best model to answer the 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? And the answer, based on this model, is yes.

As seen in Figure 8, results from the model supported our hypothesis that poor weather conditions are associated with slower average finishing times, particularly for non-elite runners. The figure below, showing feature importance by cover, suggests that average temperature, CO, dewpoint, and temp/AQI interaction are relevant drivers of an increasing average finishing time across subgroups. While they have a much lower importance score based on the model's gain values (Appendix 6, Figure 6.1), average temperature is the highest feature after those used for grouping (subgroup and gender), followed by dew point and wind speed as the most important environmental predictors. Similar results are seen in the breakdown in feature importance by frequency (Appendix 6, Figure 6.2).

Figure 8. Feature Importance by Cover from XGBoost model with engineered features



Our Predictions Answered

The predictions set at the beginning of our model inception were based on numerous research articles looking at the various weather and environmental drivers that impact a marathon runner's performance. Our model supports the findings of these articles, as well as the related predictions.

XGBoost Gain was used to see how much each feature is contributing to reducing the prediction error (XGBoost developers, 2025). Features that have a higher gain have more influence on the model, allowing us to find which features most affect finishing times and, therefore, test our predictions. We understand that gain does not prove causality directly, and it provides a useful measure of the relative importance of our features, which can be used to interpret our predictions. For more complex models, SHAP values help to show how features feed into final predictions (Awan, 2023); therefore, SHAP values were also plotted to help visualize subgroup and gender breakdowns for a few of our interactions of interest.

1. **Elite runners** will be least affected by environmental stressors due to higher fitness and experience (Ely, Martin, Cheuvront, & Montain, 2007) → *Supported*

What the model shows (Appendix 6, Figure 6.1)**:**

- Environmental variables have a very low Gain (all around 0.00002 - 0.002)
- Elite runners have much lower importance (Gain = 0.278) compared to slower runners (0.414)
- Elite runners don't have the lowest gain when compared to competitive (0.123) and recreational runners (0.075); however, this group accounts for a far smaller group size and has more extreme values.

Interpretation:

Lower importance for elite runners means that the model needs fewer splits to predict finishing times, meaning that elite performance is more consistent, less influenced by other conditions, and depends mostly on their ability and not changes in weather and air quality.

Additionally, SHAP values plotted for several environmental features by subgroup, including windspeed, average temperature, dewpoint, NO2, and CO (which are important on both the gain and cover figures) support that elite runners are one of the most resistant subgroups (Appendix 7).

2. **Competitive runners** will show greater increases in average finishing times under harsher weather and environmental conditions → *Supported*

What the model shows (Appendix 6, Figure 6.1)**:**

- subgroupcompetitive has (Gain = 0.123)
- This is:
  - lower than elites (0.278)
  - far lower than slow runners (0.414)
  - higher than recreational runners (0.075)

Interpretation:

The model splits on the competitive subgroup moderately often to reduce prediction error. This means that competitive runners have some variability in finishing times, more than recreational runners but less than slow runners. It is important to remember that the elites' subgroups' gain is likely higher due to the smaller group size and extreme finishing times, rather than environmental sensitivity.

3. **Slow runners** will experience the largest slowdowns under harsher weather and environmental conditions (Helou et al., 2012) → *-Supported*

What the model shows (Appendix 6, Figure 6.1)**:**

- subgroupslow has (Gain = 0.414), which is the largest of all features
- Gain is more than double that of any subgroup
- Gain is larger than all environmental features combined

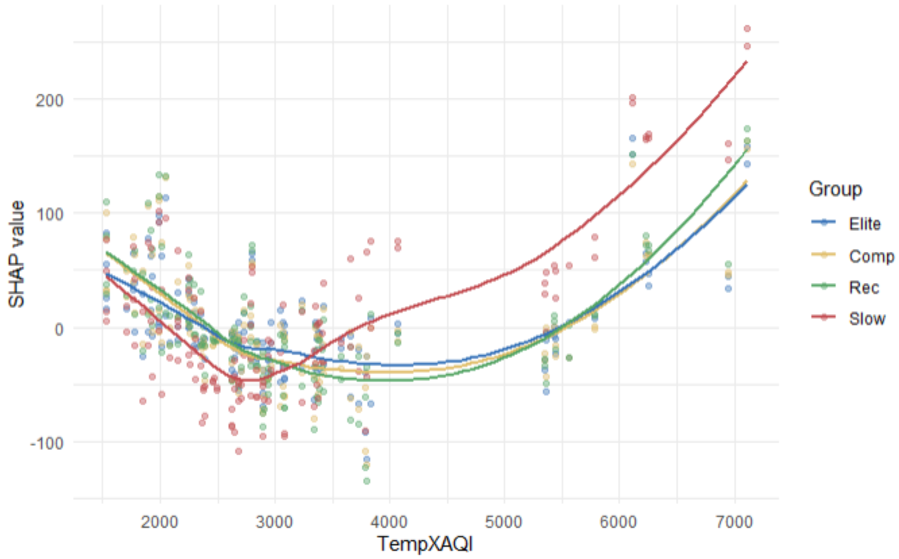
Interpretation:

This tells us that slow runners contribute to a large amount of variation in finishing times, likely meaning they predominate the dataset and are sensitive to environmental factors. We can see that the model must often split on slow runners to reduce prediction error. Figure 9 shows that the slow subgroup is significantly impacted by the temperature and AQI interaction term.

Figure 11 shows that slow runners are impacted the most by temperature changes. One reason for this may be that slower runners are exposed to the heat for longer, so the impact accumulates. Slower waves typically begin later in the day, meaning they're running during hotter conditions.

Interestingly, slow runners appear to be the most resilient to increased wind speeds (Appendix 7, Figure 7.1). This could be because higher running speeds experience greater aerodynamic drag, making wind more impactful for faster runners.

Figure 9. SHAP Values for Temperature and AQI interaction by Subgroup



4. **Female runners** will show slightly smaller performance declines under adverse weather and environmental conditions compared to males (Vihma, 2010) → *Supported*

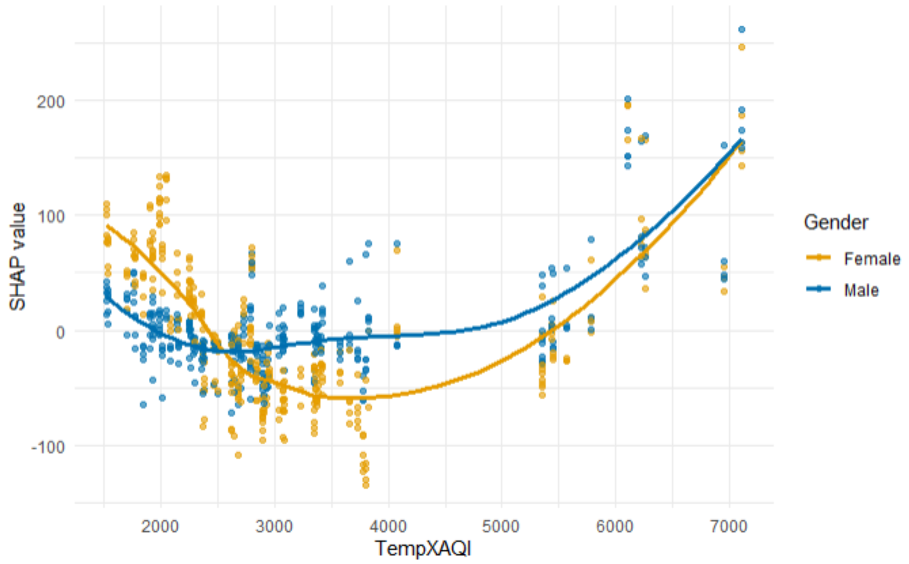
What the model shows (Appendix 6, Figure 6.1)**:**

- genderfemale has (Gain = 0.004), which is one of the highest non-subgroup features

Interpretation:

The high gain means that female finishing times have more variability, while the male times are more stable. This can be visualized in Figure 10 as female runners appear to experience a greater performance boost with optimal temperature and AQI interaction and appear more resistant than male runners to an increase in the interaction term overall. On the other hand, male runners are mostly stable with no real improvement seen, only a decrease in performance around where the interaction term reaches 5000.

Figure 10. SHAP Values for Temperature and AQI interaction by Gender



5. **Moderate temperatures and clean air** will produce the fastest finishing times across all categories (Gasparetto & Nesseler, 2020) → *Supported*

Interpretation:

The SHAP plots shown below suggest that weather and air quality shape marathon behave in a nonlinear way. Temperature shows a clear pattern, where runners are predicted to be slowest in very cold conditions, with the fastest times showing at moderate temperatures (Figure 11). Ozone follows a simpler trend, where higher levels lead to slower predicted times before the effect eventually levels off (Figure 12). PM2.5 also shows that at higher levels, there are slower predicted times (Figure 13). Overall, these plots show support for our prediction.

Figure 11. Finishing time by Average Temp and Running Subgroup

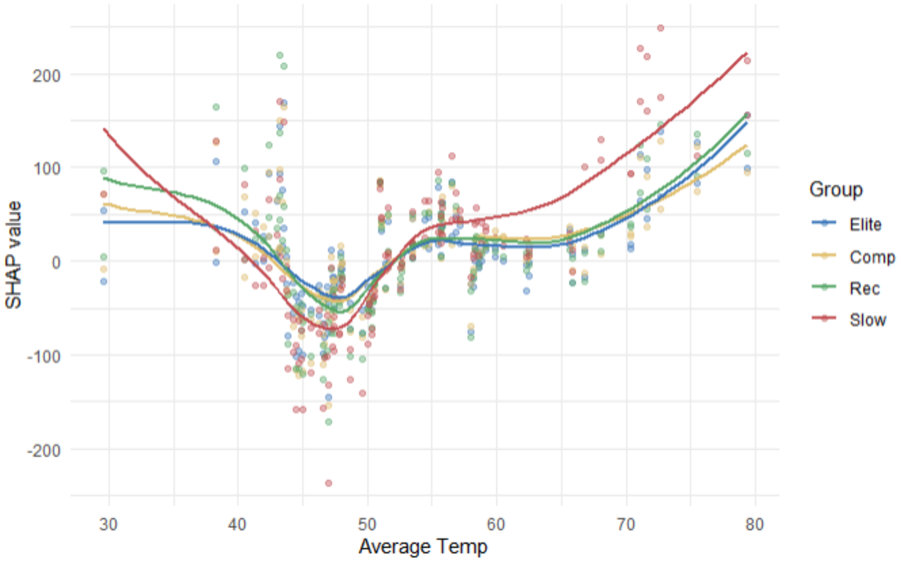


Figure 12. Finishing time by Ozone bin and Running Subgroup

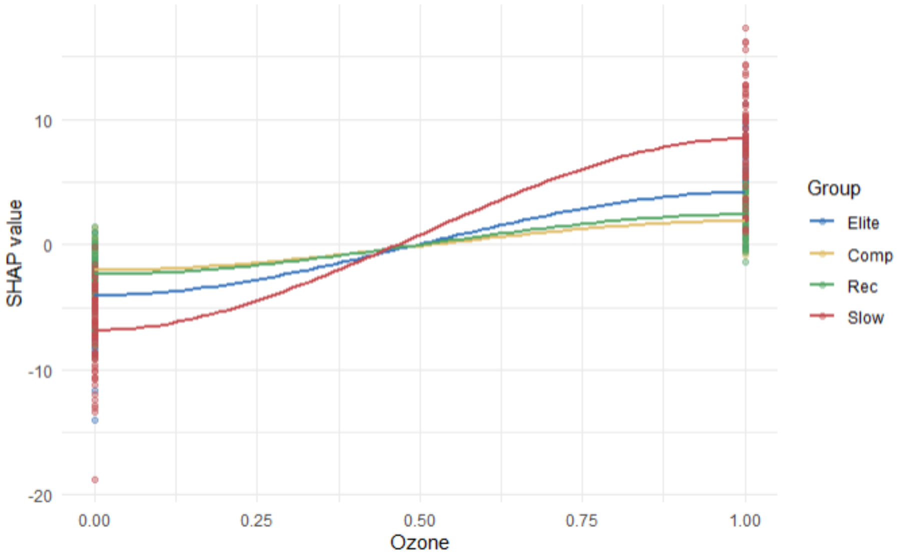
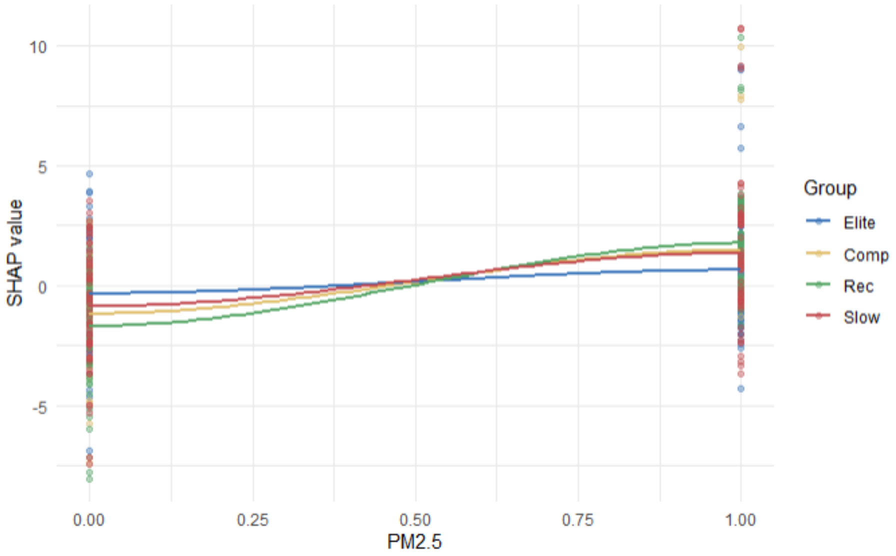


Figure 13. Finishing time by PM2.5 bin and Running Subgroup



Discussion & Next Steps

The final model has an R^2 of 0.99 on the test data, meaning that the model is able to explain > 99% of the variation in average finishing time. This model also had one of the best RMSE and MAE on the test data, and included important interaction terms and a control for supershoes (Table 5).

It should be noted that a large R^2 is likely because the highest predictors in the model are the running subgroups (Appendix 6, Figure 6.1), which were specifically engineered to group participants based on a calculated average finishing time. Future models, with more data, may benefit from creating additional interaction terms with gender and subgroups. Another option might be a model that does not aggregate the subgroups into average finishing time and instead simply adds the subgroup as a feature to the full combined race results dataset to capture the interactions. This was attempted; however, the memory required to test and train these models was not feasible with the processing power of available equipment and the timeline for this project.

As suggested above, a major limitation of this model was the availability of ‘modern’ data. This was due to time constraints and the availability of data (race, weather, and air quality) already in .csv files publicly on the internet. Also, the locations of weather stations and pollution monitors are not always close to the race course, meaning that recorded conditions may not reflect the experience of runners. Additionally, race outcomes have changed over time for reasons unrelated to environmental conditions, such as the introduction of supershoes and disruptions as a result of the COVID-19 virus. While we added controls where possible, they still introduce some uncertainty into the model.

Because the model was trained on a small sample of data, the next steps would be to bolster the data to continue to improve the RMSE and MAE, and improve the slight overfitting in the XGBoost model. Future models would benefit from more precise and contemporaneous weather and air quality measurements rather than relying on historic databases, as done in this project. Race directors can bolster this effort by incorporating their own historical race data, allowing the model to learn route-specific effects, since no two marathon courses are the same.

We see how this works as we applied the model to the held-out Berlin data with several missing features. The model was still able to make reasonably accurate predictions with an RMSE of 817 seconds (approximately 13.6 minutes) and R^2 0.93. While having all relevant environmental data would likely make the RMSE smaller, the achieved error of 13 minutes is still a practical and useful result.

Potential applications of this model by race directors include planning wave start times, optimizing water and hydration station placement, or allocating medical and security resources. Sports scientists/coaches and athletes may also benefit from this model by gaining insight into how race-day conditions may impact their performance at an upcoming race. This could be especially impactful for major qualifying races, where understanding how conditions might impact performance can inform pacing strategy, or even suggest if it is wiser to derisk injury ahead of another qualifying opportunity with more favorable race day conditions.

The model's usefulness may be further benefited by the inclusion of checkpoint times within the race to bolster medical tent, hydration station, and security planning along the race route. The model can also be further built out to support other, potentially larger participant races, including 5Ks, 10Ks, and half marathons, to support an even broader range of race directors and athletes.

While this model could use additional training, it provides a foundation for understanding the impact that weather and air quality conditions have on race performance and creates a path toward data-driven race day planning.

Code Availability

The GitHub Repo where the project lives can be found at the following website:

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

References

- Awan, A. A. (2023, June 28). *An Introduction to SHAP Values and Machine Learning Interpretability*.DataCamp.https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability
- 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.
- 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.
- Vihma, T. (2010). Effects of weather on the performance of marathon runners. International Journal of Biometeorology, 54(3), 297–306.
- XGBoost developers. (n.d.). *xgb.importance — Feature importance*. In XGBoost Documentation. Retrieved from https://xgboost.readthedocs.io/en/latest/r_docs/R-package/docs/reference/xgb.importance.html

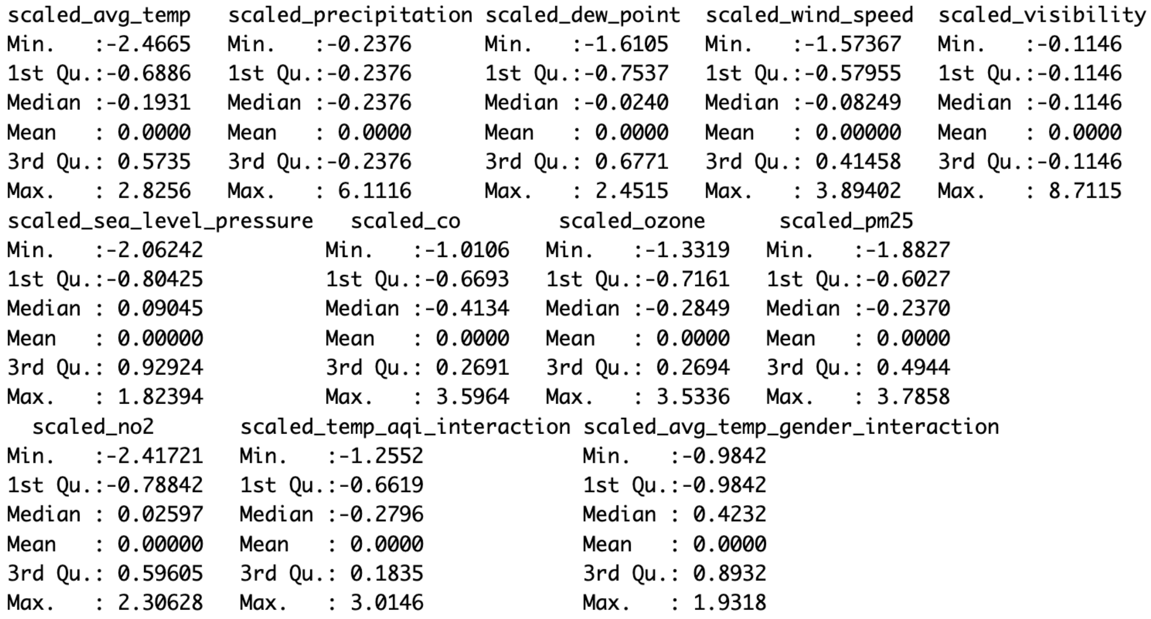
Appendix

Appendix 1. Data Dictionary

Variable	**Description**	**Type**
n	Total number of runners who completed the marathon	Integer
marathon	Name of the marathon event	Factor with 3 levels
year	Year the marathon occurred	Integer
gender	Gender category of runners (female/nonbinary and male)	Factor with 2 levels
subgroup	Running performance category: elite, competitive, average, recreational, slow	Factor with 5 levels
avg_chip_seconds	Average finishing chip time in seconds by marathon, subgroup, and gender	Numerical
avg_temp	Average race day temperature (°F)	Numerical
precipitation	Total precipitation on race day (in)	Numerical
dew_point	Average race day dew point temperature (°F)	Numerical
wind_speed	Average race day wind speed (mph)	Integer
visibility	Average race day visibility (miles)	Integer
sea_level_pressure	Average race day sea level pressure (inHg)	Numerical
co	Average carbon monoxide (AQI) on race day	Integer
ozone	Average ozone (AQI) on race day	Integer
pm25	Average PM2.5 (AQI) on race day	Integer
no2	Average nitrogen dioxide (AQI) on race day	Integer
supershoe	Indicator for supershoe prevalence (0 = low usage, 1 = high usage)	Factor with 2 levels
temp_aqi_interaction	Interaction term: avg_temp × 1(male) (0 for females, 1 for males)	Numerical

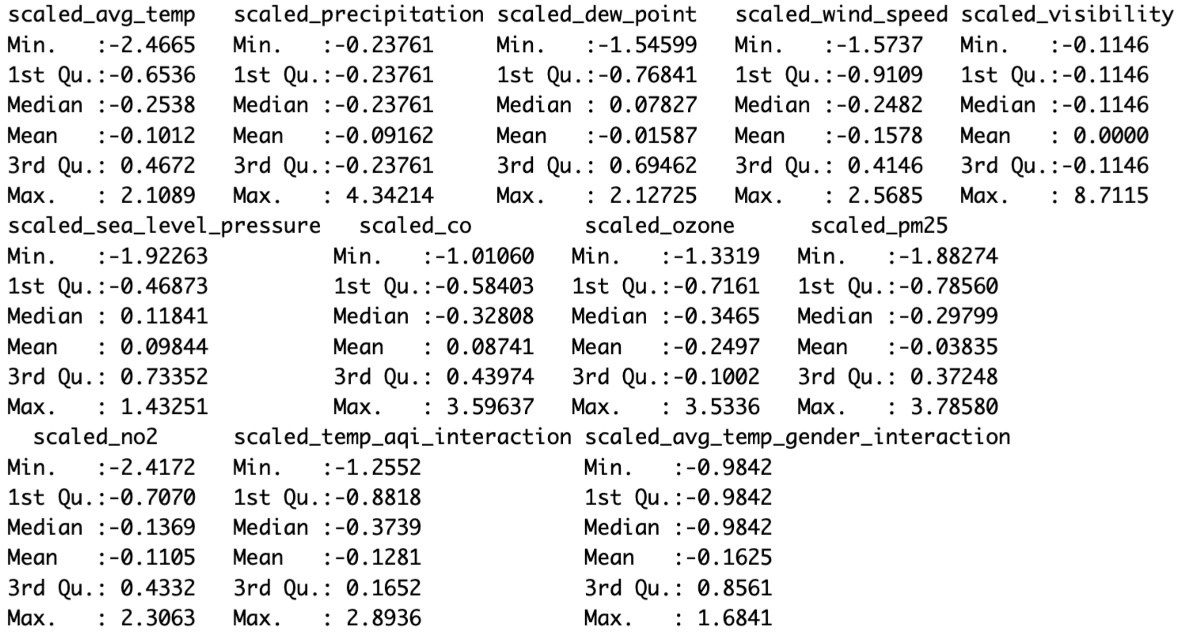
avg_temp_gender_interaction	Interaction term: avg_temp × gender (0/1-coded internally)	Numerical
ozone_bin	Binned ozone variable: 0 = low ozone, 1 = high ozone	Factor with 2 levels
pm25_bin	Binned PM2.5 variable: 0 = low PM2.5, 1 = high PM2.5	Factor with 2 levels

Appendix 2. Numeric-scaled variables for train data



```
scaled_avg_temp   scaled_precipitation scaled_dew_point  scaled_wind_speed  scaled_visibility
Min.   :-2.4665   Min.   :-0.2376      Min.   :-1.6105   Min.   :-1.57367   Min.   :-0.1146
1st Qu.:-0.6886   1st Qu.:-0.2376      1st Qu.:-0.7537   1st Qu.:-0.57955   1st Qu.:-0.1146
Median :-0.1931   Median :-0.2376      Median :-0.0240   Median :-0.08249   Median :-0.1146
Mean   : 0.0000   Mean   : 0.0000      Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
3rd Qu.: 0.5735   3rd Qu.:-0.2376      3rd Qu.: 0.6771   3rd Qu.: 0.41458   3rd Qu.:-0.1146
Max.   : 2.8256   Max.   : 6.1116      Max.   : 2.4515   Max.   : 3.89402   Max.   : 8.7115
scaled_sea_level_pressure   scaled_co        scaled_ozone      scaled_pm25
Min.   :-2.06242          Min.   :-1.0106   Min.   :-1.3319   Min.   :-1.8827
1st Qu.:-0.80425          1st Qu.:-0.6693   1st Qu.:-0.7161   1st Qu.:-0.6027
Median : 0.09045          Median :-0.4134   Median :-0.2849   Median :-0.2370
Mean   : 0.00000          Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
3rd Qu.: 0.92924          3rd Qu.: 0.2691   3rd Qu.: 0.2694   3rd Qu.: 0.4944
Max.   : 1.82394          Max.   : 3.5964   Max.   : 3.5336   Max.   : 3.7858
  scaled_no2       scaled_temp_aqi_interaction scaled_avg_temp_gender_interaction
Min.   :-2.41721   Min.   :-1.2552            Min.   :-0.9842
1st Qu.:-0.78842   1st Qu.:-0.6619            1st Qu.:-0.9842
Median : 0.02597   Median :-0.2796            Median : 0.4232
Mean   : 0.00000   Mean   : 0.0000            Mean   : 0.0000
3rd Qu.: 0.59605   3rd Qu.: 0.1835            3rd Qu.: 0.8932
Max.   : 2.30628   Max.   : 3.0146            Max.   : 1.9318
```

Appendix 3. Numeric-scaled variables for test data

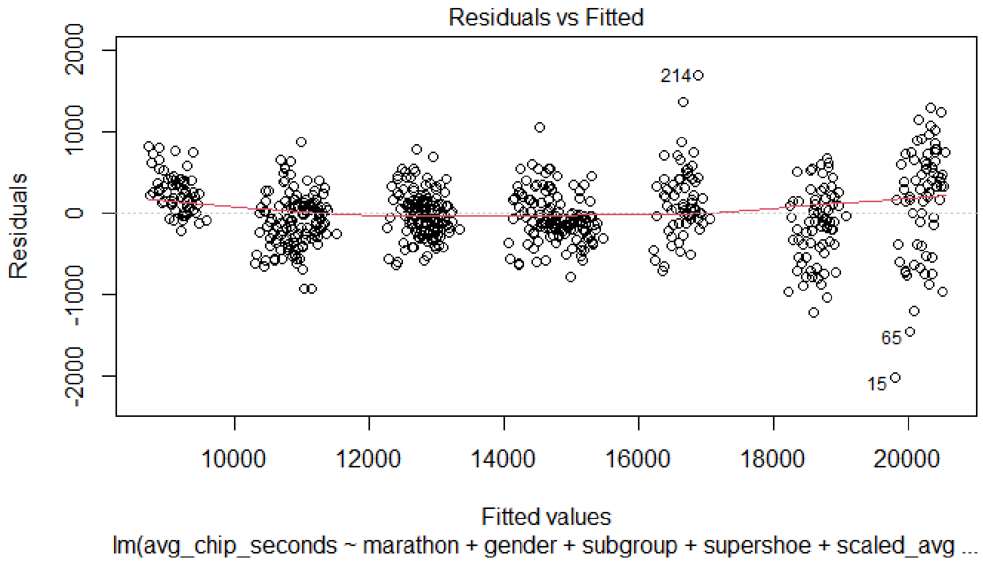


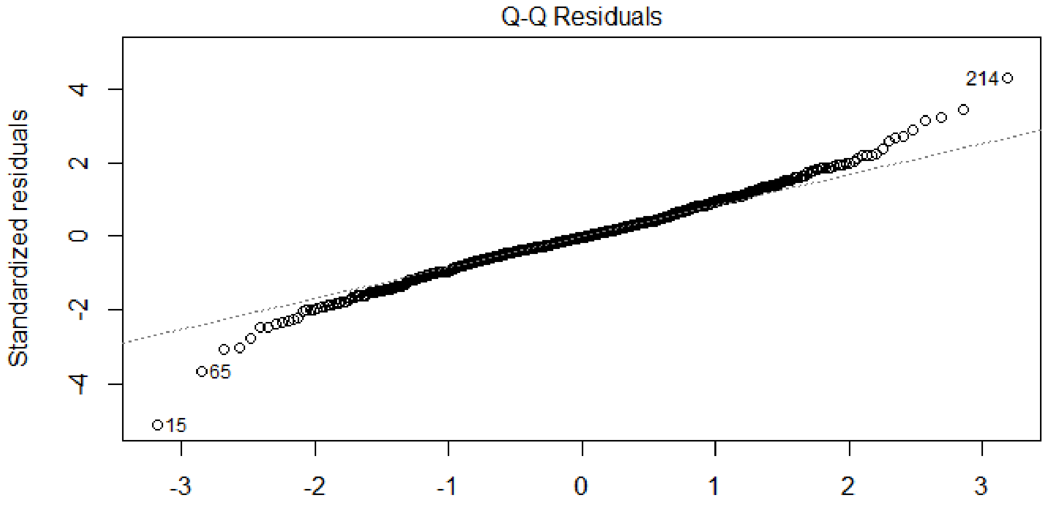
```
scaled_avg_temp   scaled_precipitation scaled_dew_point   scaled_wind_speed scaled_visibility
Min.   :-2.4665   Min.   :-0.23761     Min.   :-1.54599   Min.   :-1.5737   Min.   :-0.1146
1st Qu.:-0.6536   1st Qu.:-0.23761     1st Qu.:-0.76841   1st Qu.:-0.9109   1st Qu.:-0.1146
Median :-0.2538   Median :-0.23761     Median : 0.07827   Median :-0.2482   Median :-0.1146
Mean   :-0.1012   Mean   :-0.09162     Mean   :-0.01587   Mean   :-0.1578   Mean   : 0.0000
3rd Qu.: 0.4672   3rd Qu.:-0.23761     3rd Qu.: 0.69462   3rd Qu.: 0.4146   3rd Qu.:-0.1146
Max.   : 2.1089   Max.   : 4.34214     Max.   : 2.12725   Max.   : 2.5685   Max.   : 8.7115
scaled_sea_level_pressure   scaled_co         scaled_ozone      scaled_pm25
Min.   :-1.92263          Min.   :-1.01060   Min.   :-1.3319   Min.   :-1.88274
1st Qu.:-0.46873          1st Qu.:-0.58403   1st Qu.:-0.7161   1st Qu.:-0.78560
Median : 0.11841          Median :-0.32808   Median :-0.3465   Median :-0.29799
Mean   : 0.09844          Mean   : 0.08741   Mean   :-0.2497   Mean   :-0.03835
3rd Qu.: 0.73352          3rd Qu.: 0.43974   3rd Qu.:-0.1002   3rd Qu.: 0.37248
Max.   : 1.43251          Max.   : 3.59637   Max.   : 3.5336   Max.   : 3.78580
  scaled_no2      scaled_temp_aqi_interaction scaled_avg_temp_gender_interaction
Min.   :-2.4172   Min.   :-1.2552            Min.   :-0.9842
1st Qu.:-0.7070   1st Qu.:-0.8818            1st Qu.:-0.9842
Median :-0.1369   Median :-0.3739            Median :-0.9842
Mean   :-0.1105   Mean   :-0.1281            Mean   :-0.1625
3rd Qu.: 0.4332   3rd Qu.: 0.1652            3rd Qu.: 0.8561
Max.   : 2.3063   Max.   : 2.8936            Max.   : 1.6841
```

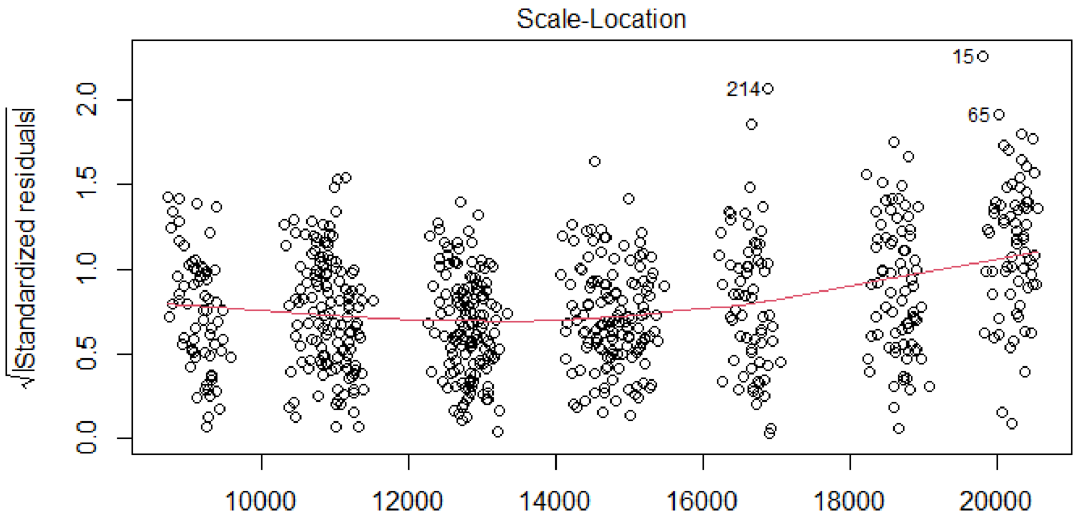
Appendix 4. Table of Feature Engineering

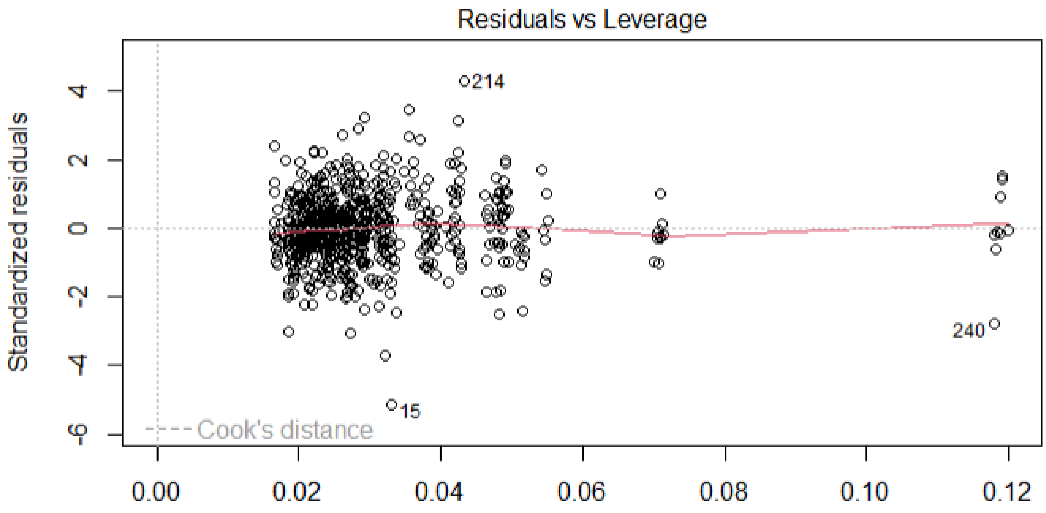
Feature Name	**Description**
supershoe	Binary control variable: 1 if year ≥ 2018 (to account for use of advanced running shoes), 0 otherwise
temp_dew_interaction	Interaction of average temperature and dew point
temp_aqi_interaction	Interaction of average temperature and AQI
temp_precip_interaction	Interaction of average temperature and precipitation
temp_wind_interaction	Interaction of average temperature and wind speed
pm25_temp_interaction	Interaction of PM2.5 and average temperature
dew_wind_interaction	Interaction of dew point and wind speed
pressure_temp_interaction	Interaction of sea level pressure and average temperature
avg_temp_gender_interaction	Interaction of average temperature and gender (male coded as 1, female as 0)

Appendix 5. Linear Model Assumptions






Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
10000
12000
14000
16000
18000
20000
15
214
65


Residuals vs Leverage
Standardized residuals
-6
-4
-2
0
2
4
0.00
0.02
0.04
0.06
0.08
0.10
0.12
214
15
240
Cook's distance

Appendix 6. Feature Importance Results of XGBoost Model with raw features

Figure 6.1 Feature Importance by Gain

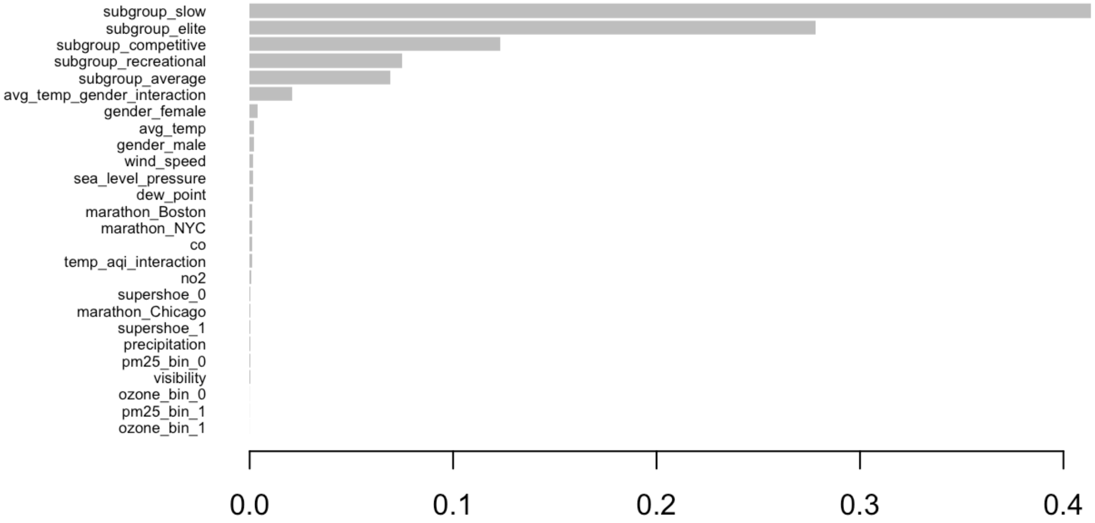
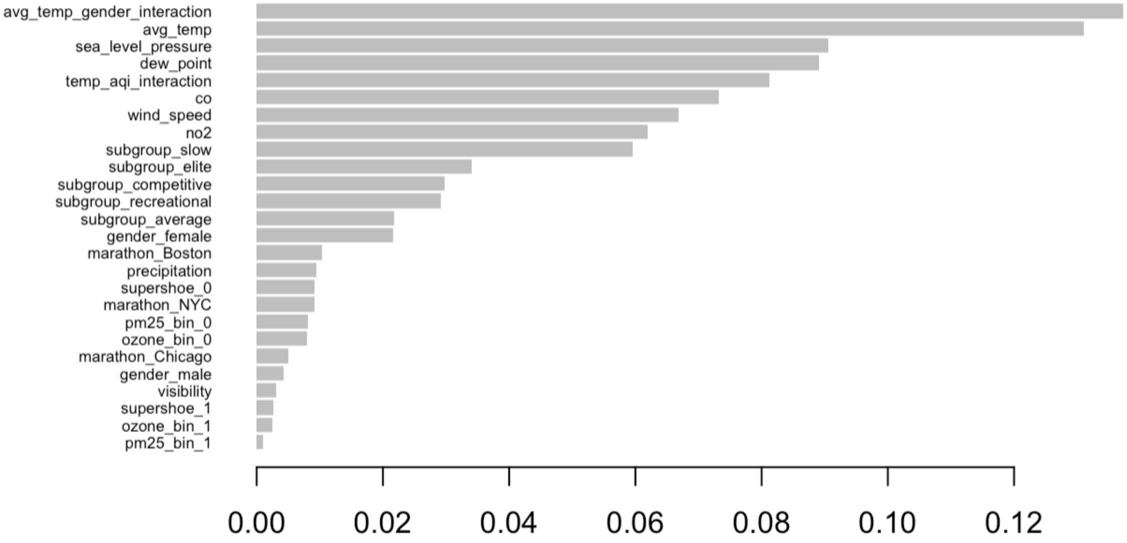


Figure 6.2 Feature Importance by Frequency



Appendix 7. SHAP Figures

Figure 7.1 SHAP Values for Windspeed by Subgroup

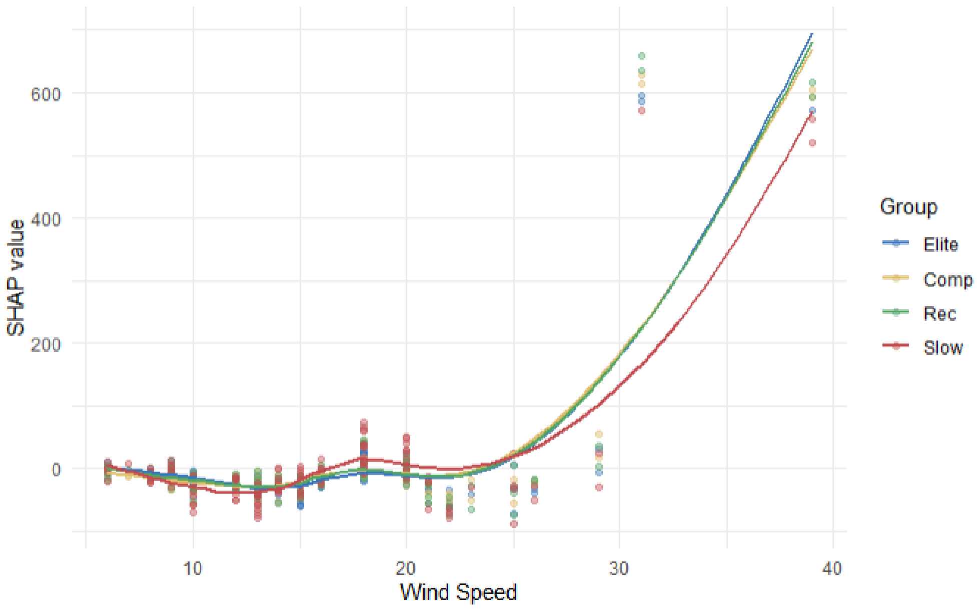


Figure 7.2 SHAP Values for Dewpoint by Subgroup

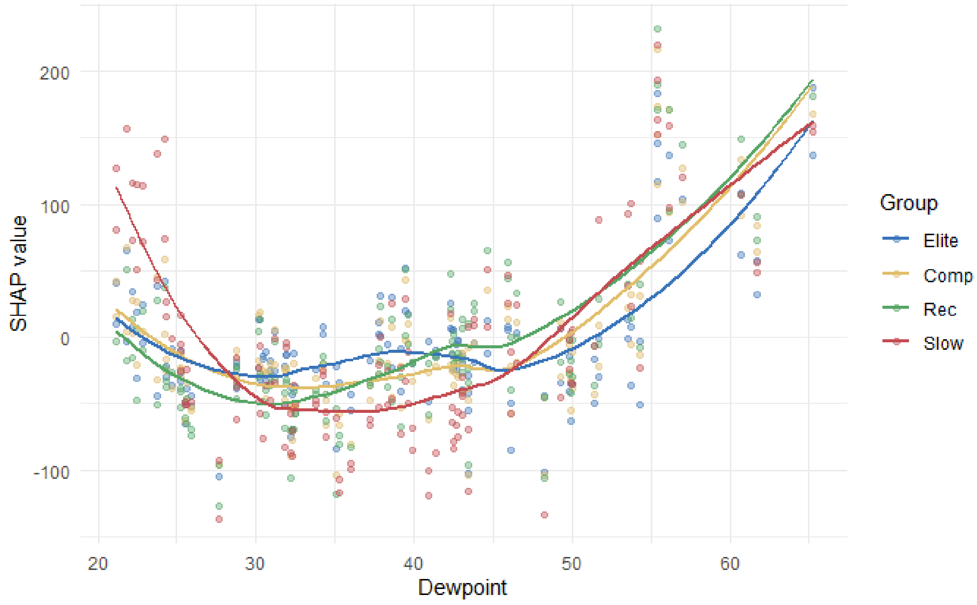


Figure 7.3 SHAP Values for NO2 by Subgroup

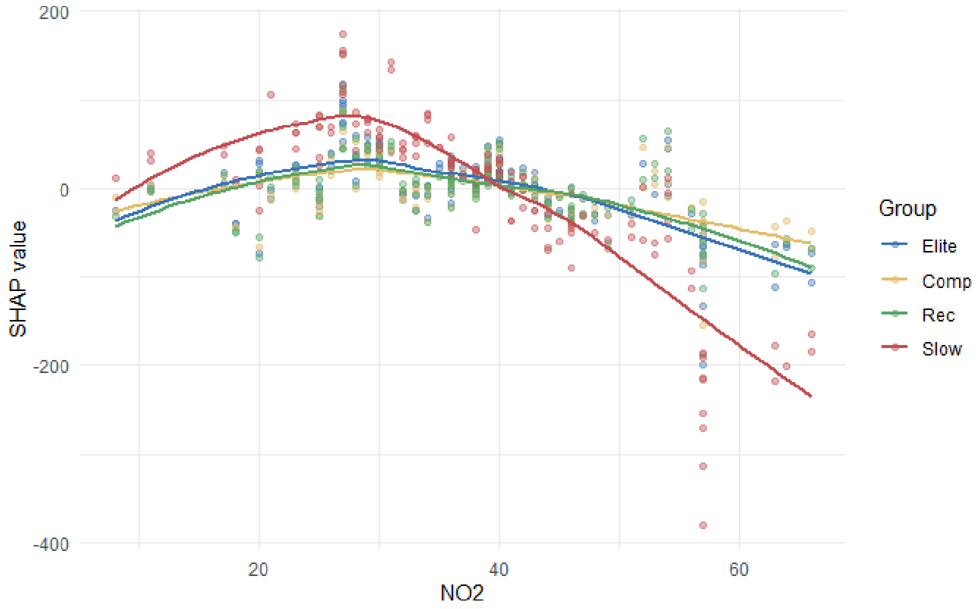
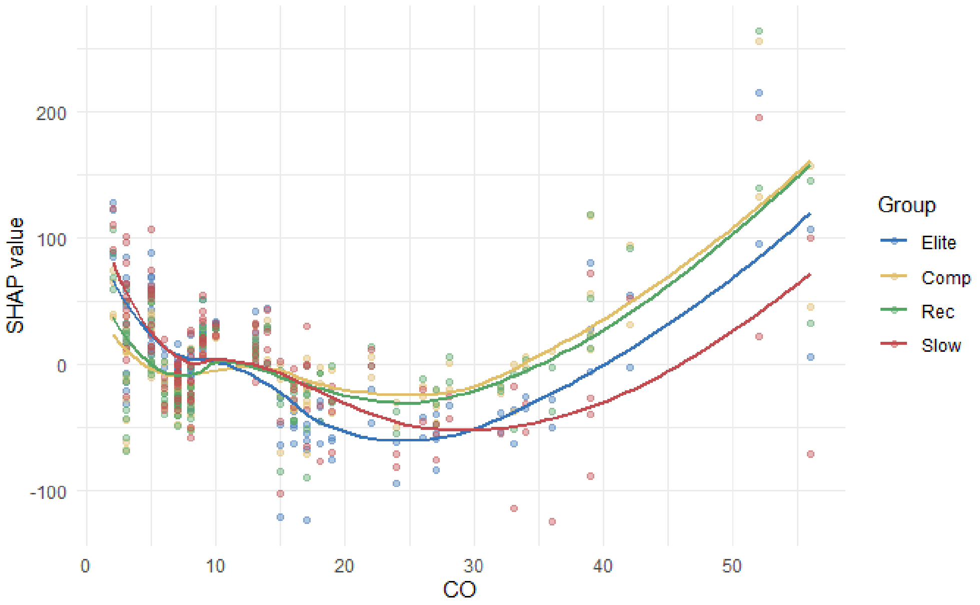


Figure 7.4 SHAP Values for CO by Subgroup


SHAP value
CO
Group
Elite
Comp
Rec
Slow