Combined Report:

Preprocessing & Initial Model

Predicting Marathon Average Finishing Times: The Impact of Weather and Air Quality on Marathons

Team Lead: Meghan Holden

Recorder: Krisha Bugajski-Sharp

Spokesperson: Zachary D'Urso

Link to GitHub:

BugajskiSharp/capstone-project-team-c: Team C data capstone project.

Table of Contents

Introduction

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. Wouldn't it be helpful if we could predict how well runners will perform based on these factors?

In this project, we will analyze major marathons, including Boston, Chicago, and New York, to understand how race day weather and air quality influence average finishing times. We will start by combining marathon results with historical EPA and Weather Underground data, cleaning and pre-processing the variables, handling missing data, and creating new features that can capture the combined effects of these environmental variables. Because our early exploratory work showed that percentile-based groups can be misleading, we instead define runner performance subgroups using each runner's ratio to the winner's time, giving more consistent comparison across years and marathons.

Our goal is to build and compare several models. We will use supervised approaches such as versions using decision tree-based binning and a regularised LASSO model to evaluate whether binned features can improve predictive performance. We will then make a Multiple Linear Regression model to see fundamental relationships, followed by a Gradient Boosting model, such as XGBoost or LightGBM, to find which environmental factors matter the most. Overall, we want to understand which conditions slow runners

down and which modeling choices produce the most accurate and interpretable predictions.

Background & 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, Chicago, and Berlin.

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.

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).

Pre-Processing and Feature Engineering Methods

Our initial preprocessing and feature engineering plan involved standardizing datasets across marathons, handling missing values, scaling continuous variables, and engineering interaction terms and a control variable for supershoes. Exploratory data analysis showed that pollution and temperature were associated with slower marathon performance across subgroups (Appendix 2 and Appendix 3). It also showed that our quintile subgrouping produced misleading yearly patterns because slower participation increased over time (Appendix 4). EDA also showed substantial missing environmental data for Berlin, so it was separated into a case study rather than included in the main modeling dataset.

The initial model will use our previously engineered features along with the ozone and PM2.5 bins derived from the decision tree in the supervised feature engineering section.

We will fit multiple linear regression models using different combinations of these binned and continuous predictors to compare performance and to determine which combination of the environmental variables should be used in the final modeling stage.

Pre-Processing & Feature Selection

We started by merging several marathon datasets together, and then merging them with corresponding weather and air quality datasets, making sure there were consistent naming, date formats, and finishing time conversions. Then variable types were checked to confirm that continuous predictors were stored as numeric values and categorical identifiers, such as marathon, gender, and subgroup, were stored as factors so the models could interpret them correctly. Because the Berlin marathon had a substantial amount of missing values for CO, PM10, PM2.5, and visibility, we removed Berlin from the primary modeling dataset and assigned it as a separate case study. For the remaining marathons (Chicago, Boston, and NYC), we evaluated missingness patterns to determine whether imputation or variable removal was more appropriate.

Next, we examined histograms for all continuous variables to check distribution shapes and also to verify the accuracy of finishing time conversions (Appendix 6). Several variables showed right-skewed distributions (e.g. PM2.5 and wind speed), but no transformation was required for our intended modeling approaches. After preprocessing, we made a 90–10% training–test split. We chose this ratio because removing Berlin reduced our dataset size, and a larger training set helps improve model stability while still giving a meaningful evaluation. Since we are fitting multiple linear regression and gradient boosting models, which perform best when numeric predictors are on similar

scales, we decided to scale all continuous predictors except the outcome (avg_chip_seconds) and identifiers, such as year and number of runners (n), on both the training and test data.

Feature Engineering Methods

Our original plan included constructing several interaction terms (e.g. temperature × dew point, AQI × temperature), creating supershoe and COVID-19 control variables, and finalizing performance subgroup definitions. EDA also showed that quintile-based subgroups increased over time due to growth in slower participants. We therefore replaced the quintile method with performance groups defined by each runner's ratio to the winner's time, which allowed for consistent subgroup definitions across years, marathons, and genders and avoided any distortions caused by participation trends (Appendix 4). We used a histogram that showed a naturally right-skewed distribution with clear clustering to help guide the data-driven performance cutoffs and explained why slower groups dominated the counts, while also confirming that both genders (where females were grouped with nonbinary) share similar patterns and supported the use of ratio-based performance categories.

We also evaluated multicollinearity by computing a correlation matrix that included both original and engineered variables. Earlier, EDA showed that *AQI* was highly correlated with *PM2.5*, so *AQI* was removed to avoid redundancy. We also removed *main_pollutant* because it is essentially the categorical counterpart of *AQI* and provided no additional independent information.

Unsupervised Feature Engineering

We considered Principal Component Analysis (PCA) as an unsupervised feature-engineering method to reduce dimensionality and address multicollinearity. However, PCA was found to be inappropriate for our data for several reasons.

First, our predictors, such as temperature, dew point, wind speed, PM2.5, precipitation, and sea level pressure, carry clear physical and environmental meaning. If we used PCA, it would combine our meaningful environmental variables into new components that are difficult to interpret in terms of 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.

We also considered using k-means clustering as an unsupervised method. However, it did not make sense for our data and therefore was not used. The performance groups are based on clear ratios rather than hidden patterns in the data, so clustering would not align with categories established for elite, competitive, average, recreational, and slow runners. Also, the environmental variables change gradually, and do not fall into separate categories, meaning that using k-means clustering would not produce meaningful groups.

Supervised Feature Engineering

Decision-Tree Feature Selection

Tree-based methods were used to determine whether certain continuous variables are better left as continuous or grouped into categories based on meaningful thresholds. All features from feature engineering thus far were included in the decision tree model to

investigate potential thresholds. These features include: avg_temp, precipitation, dew_point, wind_speed, visibility, sea_level_pressure, co, ozone, pm25, no2, supershoe, temp_aqi_interaction, and covid_era. Performance subsets and gender were not included in this feature engineering investigation, as the intention of this model is to investigate potential binning of continuous predictors, not which features predict average finishing time overall.

This was done first by creating a dataset averaging overall chip time by marathon, by year, removing Berlin data due to the missing values, and then running a decision tree with the training data that came from the split of 90-10% with the features described above to predict avg_chip_seconds.

LASSO Feature Selection

Next, we ran a LASSO regression to see whether using binned air quality features improves predictive performance. We intentionally decided to use only ozone and PM2.5 (continuous + binned versions) in the LASSO model. We decided that including all predictors would overwhelm the model and prevent us from isolating the effect of binned vs continuous variables.

Initial Model Methods

Based on the binning recommendations from the decision tree and LASSO regression above, we are running two multiple linear regression models as our initial model, testing the recommended bins against their continuous counterparts. Multiple linear regression

models allow us to understand how each predictor impacts the outcome and provide a solid baseline for continued training and model improvement.

Model Assumptions

The multiple linear regression model makes several assumptions about the data. *ISLR* (2021) explains these model assumptions as:

1. A linear relationship between the predictors and the response
2. Error terms are uncorrelated
3. Error terms have constant variance (also known as homoscedasticity)
4. No high-leverage points
5. No high collinearity between features

Model Methods

Two multiple linear regression models will be run, testing feature combinations of PM2.5 in its binned or scaled form (based on results from the decision tree and LASSO described below) using the lm() function in R. Each model will have model assumptions assessed, using the plot() function in R, to ensure the data is appropriate for the model.

The results held in the summary() function will be compared across each model, looking at coefficient p-values and f-statistics to determine the relationship between predictors and outcome in each model. R^2 and residual standard error (RSE) will be compared to determine how well each model fits the data.

Then, the models will be tested against the held-out test data set to check and compare model performance.

Results

Pre-processing & Feature Engineering Results

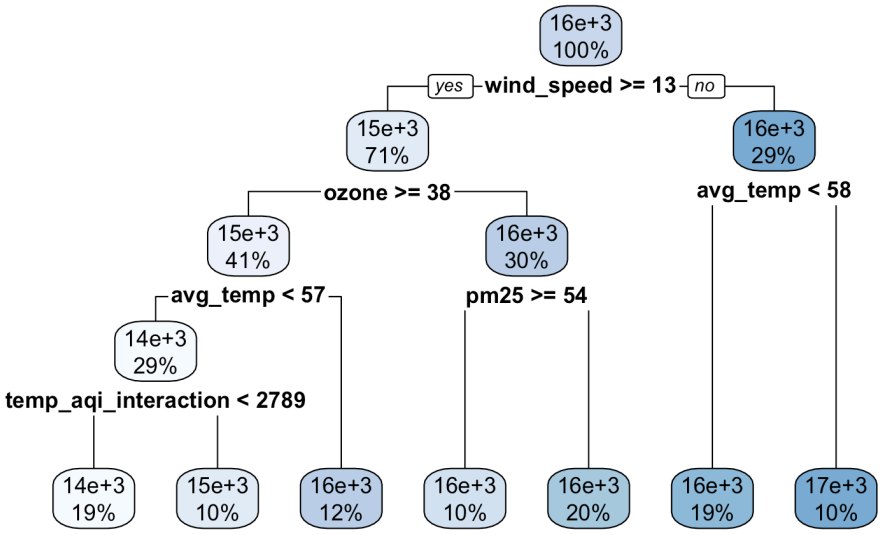
After separating Berlin due to extensive missing environmental data, our main dataset contained 770 observations and 21 variables. We assessed missingness and removed PM10 because it had 320 missing values and would introduce noise rather than signal, while PM2.5, which is an important pollutant with only 70 missing values, was retained and imputed using KNN (k = 5) to preserve structure and variability without getting rid of any rows. We then evaluated continuous variable histograms to confirm distribution shapes and verify finishing time conversions; several predictors, including PM2.5 and wind speed, displayed expected right skew, but no transformations were required because of our modeling goals (Appendix 6). Performance subgroups were reconstructed using runner-to-winner time ratios, which helped to eliminate the misleading yearly patterns made by the quintiles and helped to align with natural clustering observed as seen in Appendix 5. After confirming all variable types, we added interaction terms and the supershoe control (Appendix 7), then used a correlation matrix to identify multicollinearity and removed highly correlated engineered features (Appendix 8), saving only meaningful and non-redundant predictors such as *temp_aqi_interaction*, *supershoe*, and *avg_temp_gender_interaction* (Appendix 9). We then made a 90–10% training–test split to maximize training stability, given the reduced dataset, and all the numeric predictors, not including avg_chip_seconds and identifiers such as *year* and *n*, were

scaled separately for the training-test datasets to prevent any data leakage (Appendix 10 and Appendix 11).

Then we used the Decision Tree to find meaningful binned features. Figure 1 shows important feature thresholds when wind speed is greater than or equal to 13, ozone is greater than or equal to 38, PM2.5 is greater than or equal to 54, average temperature is less than 57, and temperature and AQI interaction is less than 2789.

The tree suggests that wind speed, ozone, and average temperature are the strongest predictor variables, and PM2.5 and temperature and AQI interaction. However, to be thorough, several models were compared, looking at these features as both their original continuous values and their binned values based on the decision tree.

Figure 1. Decision Tree Results of Original Continuous Features for Predicting avg_chip_seconds using training data only.



Several linear regression models were compared with both binned and continuous variables based on the decision tree to help guide which features might most benefit the

model through binning, versus remaining continuous. The binned variables investigated are shown in Table 1.

Table 1. Binned Features to Investigate

Feature Name	**Binning Description**
wind_bin	High: ≥ 13 Low: < 13
temp_bin	Cool/Moderate: ≤ 58 Warm: > 58
ozone_bin	High: ≥ 38 Low: < 38
pm25_bin	High: ≥ 54 Low: < 54
temp_aqi_bin	High interaction: ≥ 2789 Low interaction: < 2789

Table 2 provides a breakdown of the various models comparing results of different bins to see which combination most improves their mean squared error (RMSE), R^2, and mean absolute error (MAE).

Table 2. Comparison of Models with Binned Features

Model Name	**Model features**	**Model Results**
Continuous (original)	*All continuous:* avg_temp, ozone, pm25, wind_speed, temp_aqi_interaction	RMSE: 895.8262 R-squared: 0.4873768 MAE: 751.7258
Binned	*All binned:* temp_bin, ozone_bin, pm25_bin, wind_bin, temp_aqi_bin	RMSE: 868.4068 R-squared:0.5071339 MAE: 709.1803
Temp Binned	*Binned:* temp_bin *Continuous:* ozone, pm25, wind_speed, temp_aqi_interaction	RMSE: 896.0868 R-squared: 0.4642291 MAE: 751.9601

Temp & Ozone Binned	*Binned:* temp_bin, and ozone_bin *Continuous:* pm25, wind_speed, temp_aqi_interaction	RMSE: 857.2706 R-squared: 0.5004747 MAE: 706.8899
Ozone & PM2.5 Binned	*Binned:* ozone_bin and pm25_bin *Continuous:* avg_temp, wind_speed, temp_aqi_interaction	RMSE: 763.2415 R-squared: 0.5920959 MAE: 635.5167
Ozone Binned	*Binned:* ozone_bin *Continuous:* avg_temp, pm25, wind_speed, temp_aqi_interaction	RMSE: 802.1434 R-squared: 0.622678 MAE: 672.732
Wind Binned	*Binned:* wind_bin *Continuous:* avg_temp, pm25, wind_speed, temp_aqi_interaction	RMSE: 930.0377 R-squared: 0.4669259 MAE: 777.1032
Wind, Ozone & PM2.5 Binned	*Binned:* ozone_bin, wind_bin, pm25_bin *Continuous:* avg_temp, temp_aqi_interaction	RMSE: 805.9406 R-squared: 0.6037733 MAE: 671.5632

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.

Based on these results, we will incorporate the binned ozone and PM2.5 features into our final dataset for modeling and test a couple of initial models with the continuous version of these features, as well as the binned version.

Then, LASSO regression was applied for further feature selection. The results showed that the LASSO retained both the binned versions of pm2.5 and ozone, as well as the scaled_pm2.5 (Table 3). However, the continuous scaled_ozone was removed. This means that the binned air quality values and continuous scaled_pm2.5 all carry the

strongest predictive signals for average chip time, and the continuous scaled_ozone variable does not add meaningful predictive value. Table 4 shows the key metrics from the LASSO model. Because the predictors can only explain a small portion of finishing times, it was expected that the model's errors would be large. The selected lambda (0.0244) shows that only a mild penalty was needed. Although the RMSE and MAE are high, this makes sense for the reduced model that excludes the major drivers of runner performance. Overall, LASSO helped confirm which environmental variables carry the strongest signals, helping guide what we should include in the full predictive model.

Table 3. LASSO model results, where (.) means the variable was retained and 0 means the variable was shrunk to zero and not retained.

Feature	**Coefficient**
Intercept	14227.13
scaled_ozone	0.00
scaled_pm25	.
ozone_bin	.
ozone_pm25	.

Table 4. Key metrics from the LASSO model

Metric	**Value**
Lambda Selected	0.0244
Intercept	14,227 seconds
RMSE	3,383
MAE	2,861

Overall, the two supervised feature selection steps confirm that binning PM2.5 and ozone improves the model's predictive performance.

Initial Model Results: Algorithm, Assumptions, Overfitting

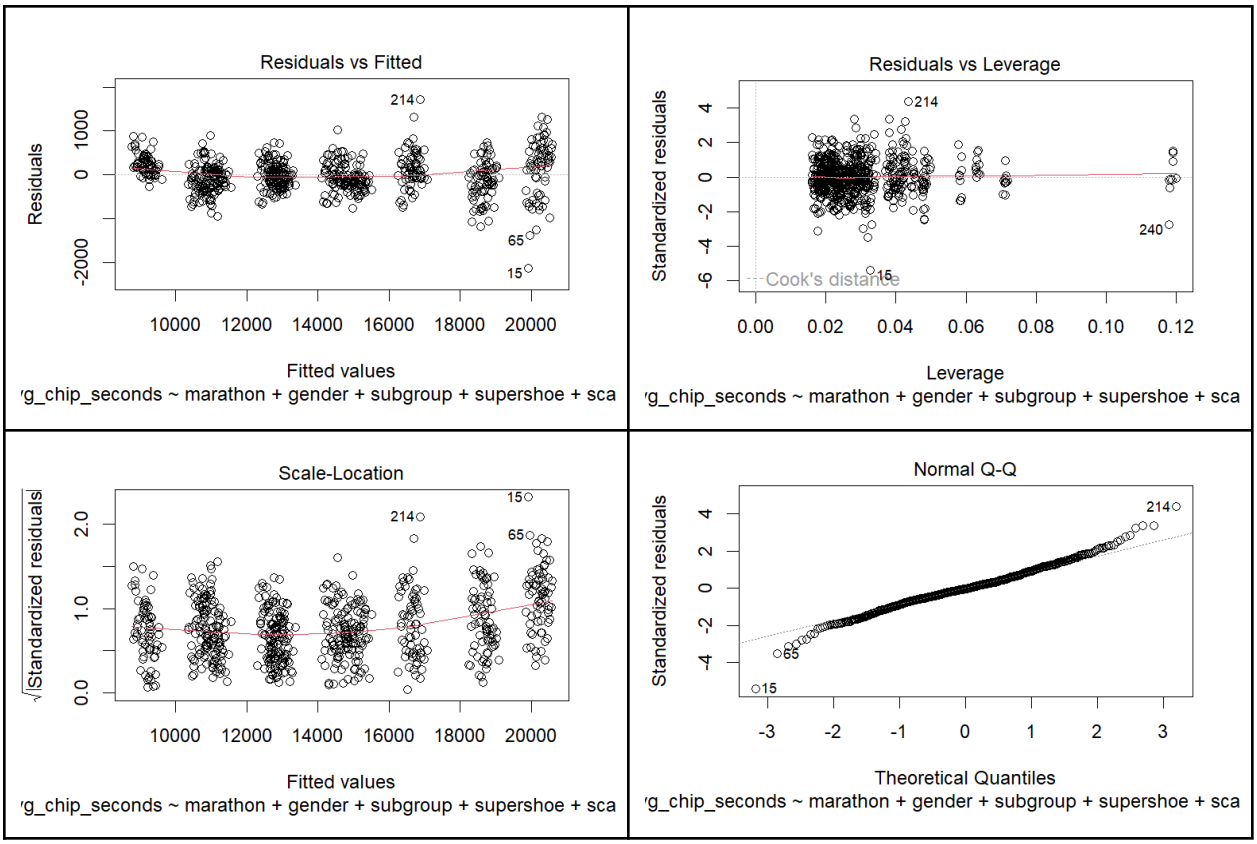
Two models were tested and compared, differing only by the feature scaled_pm25 and pm25_bin (the continuous scaled and the binned version of the PM2.5 feature). The initial multiple linear regression model, shown below in Function 1, was run with all weather and environmental features scaled (except for ozone, which is binned in both models based on LASSO results). Note that COVID-19 was not included in the initial models, as there was minimal data coming from marathons after 2020 in the dataset, this may be investigated in further models.

Function 1. Initial Model with ozone_bin and all scaled features

```
lm(formula = avg_chip_seconds ~ marathon + gender + subgroup +
   supershoe + scaled_avg_temp + scaled_precipitation + scaled_dew_point +
   scaled_wind_speed + scaled_visibility + scaled_sea_level_pressure +
   scaled_co + scaled_pm25 + ozone_bin + scaled_no2 + scaled_temp_aqi_interaction +
   scaled_avg_temp_gender_interaction, data = train_data)
```

The assumptions for multiple linear regression are satisfied based on the results in Figure 2. The Residual vs Fitted plot shows the residuals are mostly scattered around the horizontal line with no clear pattern satisfying the linearity assumption. There might be slight funneling in the Residuals vs Fitted and Scale-Location, suggesting homoscedasticity is mostly maintained. There are two high leverage points in the Residuals v Leverage plot (214 and 240), suggesting potential influence from these values. Future models may be tested to account for these influential points. The Q-Q plot shows that the residuals mostly follow the diagonal line, suggesting normality of residuals.

Figure 2. Model Assumptions for Initial Model with ozone_bin and all scaled features



Results of the model are shown in Table 5. The model has high R2 and adjusted R2 (0.9866 and 0.9862, respectively), suggesting it is a good fit for the data. The results of the model are highly significant as shown by the F-statistic (2481) and p-value (< 2.2e-16).

As expected, supershoes have an impact on the finishing time, making runners 163.51 seconds (~2.7 minutes) faster. Looking at the environmental and weather features, the only features that are not significant are scaled_precipitation, scaled_co, and scaled_avg_temp_gender_interaction. Dewpoint, windspeed, visibility, sea level pressure, high ozone, and AQI-temperature interaction all contribute to slower speeds.

The results in this model of average temperature, PM2.5, and NO2 leading to faster times is counterintuitive and may need to be further investigated.

Table 5. Model Summary for Initial Model with ozone_bin and all scaled features

```
Residuals:
     Min       1Q   Median       3Q      Max
-2129.62  -223.98   -11.79   231.92  1714.68

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       14127.20     115.78 122.018  < 2e-16 ***
marathonChicago                     403.02      82.59   4.880 1.33e-06 ***
marathonNYC                         622.45      67.10   9.276  < 2e-16 ***
gendermale                        -1676.30     173.95  -9.637  < 2e-16 ***
subgroupcompetitive               -1809.36      48.49 -37.317  < 2e-16 ***
subgroupelite                     -3762.47      48.25 -77.981  < 2e-16 ***
subgrouprecreational               2140.38      48.14  44.460  < 2e-16 ***
subgroupslow                       5730.84      47.62 120.351  < 2e-16 ***
supershoe1                         -163.51      45.60  -3.586 0.000360 ***
scaled_avg_temp                    -173.19      52.49  -3.299 0.001021 **
scaled_precipitation                 10.40      18.02   0.577 0.563929
scaled_dew_point                    122.22      31.65   3.861 0.000124 ***
scaled_wind_speed                    95.19      21.52   4.424 1.13e-05 ***
scaled_visibility                    86.11      17.43   4.939 9.90e-07 ***
scaled_sea_level_pressure           111.50      28.73   3.881 0.000114 ***
scaled_co                           -36.12      23.46  -1.540 0.124097
scaled_pm25                        -104.28      48.32  -2.158 0.031264 *
ozone_bin1                          197.37      50.92   3.876 0.000116 ***
scaled_no2                          -68.94      24.80  -2.780 0.005594 **
scaled_temp_aqi_interaction         226.20      62.54   3.617 0.000320 ***
scaled_avg_temp_gender_interaction   49.96      88.29   0.566 0.571682
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 400.8 on 672 degrees of freedom
Multiple R-squared:  0.9866,   Adjusted R-squared:  0.9862
F-statistic:  2481 on 20 and 672 DF,  p-value: < 2.2e-16
```

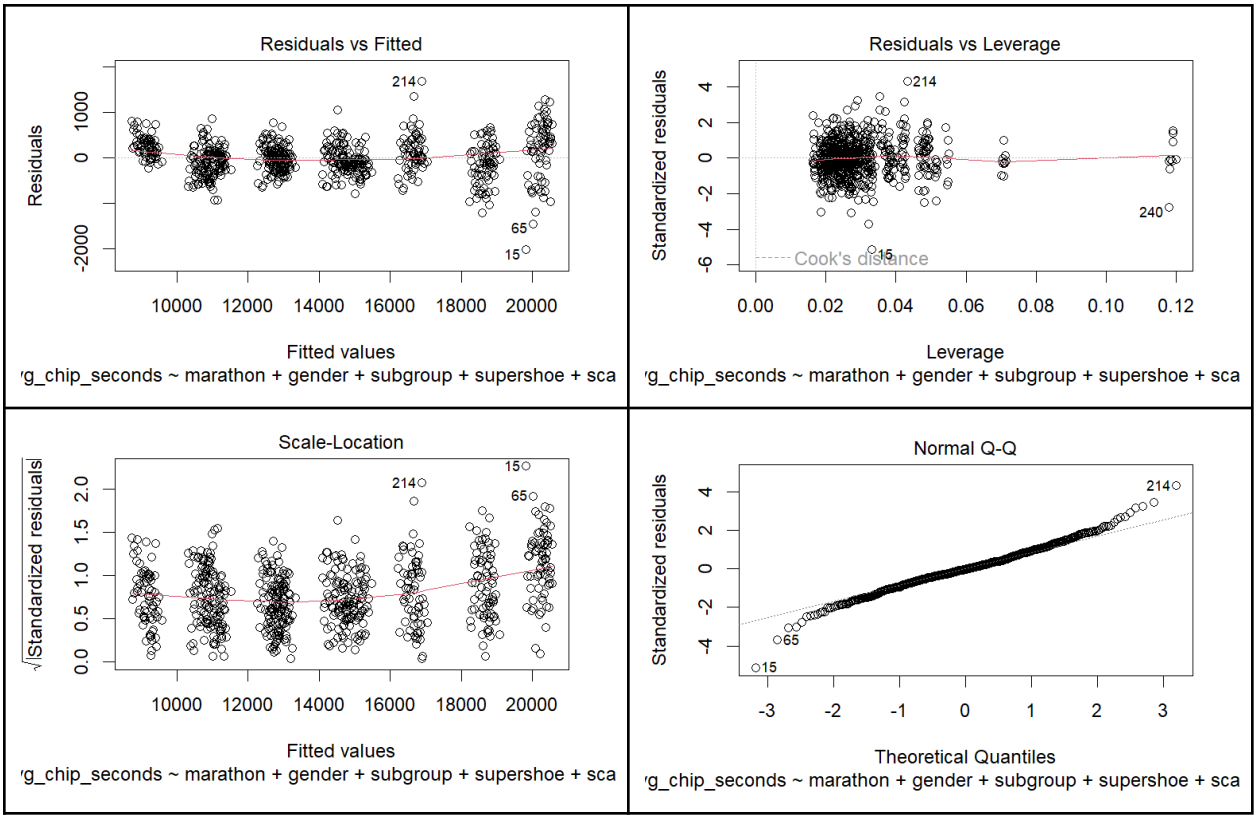
Function 2 shows nearly an identical model, with the only difference being the substitution of PM2.5 as a binned feature instead of a continuous feature.

Function 2. Initial Model with ozone_bin, pm25_bin, and all scaled features

```
lm(formula = avg_chip_seconds ~ marathon + gender + subgroup +
    supershoe + scaled_avg_temp + scaled_precipitation + scaled_dew_point +
    scaled_wind_speed + scaled_visibility + scaled_sea_level_pressure +
    scaled_co + pm25_bin + ozone_bin + scaled_no2 + scaled_temp_aqi_interaction +
    scaled_avg_temp_gender_interaction, data = train_data)
```

The plots in Figure 3 are very similar to those for the previous model, suggesting a similar interpretation of assumptions.

Figure 3. Model Assumptions for Initial Model with ozone_bin and all scaled features



The results of the model shown in Table 6 are very similar to the previous model. The model has identical R2 and adjusted R2 (0.9866 and 0.9862, respectively), suggesting it is a good fit for the data. The results of the model are highly significant as shown by the F-statistic (2478) and p-value (< 2.2e-16).

The major difference between the model with PM2.5 binned versus the model with PM2.5 continuous is that PM2.5, as a binned feature, shows a slowing effect to runners versus an enhancement when it is in the model as a continuous feature.

Table 6. Model Summary for Initial Model with ozone_bin, pm25_bin, and all scaled features

```
Residuals:
     Min       1Q   Median       3Q      Max
-2021.98  -222.10    -9.13   227.41  1685.71

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       14111.65     117.88 119.711  < 2e-16 ***
marathonChicago                     374.67      79.99   4.684 3.41e-06 ***
marathonNYC                         609.27      66.09   9.219  < 2e-16 ***
gendermale                        -1690.84     174.11  -9.711  < 2e-16 ***
subgroupcompetitive               -1812.13      48.51 -37.356  < 2e-16 ***
subgroupelite                     -3763.43      48.28 -77.957  < 2e-16 ***
subgrouprecreational               2138.29      48.18  44.384  < 2e-16 ***
subgroupslow                       5728.90      47.64 120.251  < 2e-16 ***
supershoe1                         -155.72      45.82  -3.399 0.000716 ***
scaled_avg_temp                     -82.71      40.71  -2.032 0.042586 *
scaled_precipitation                 22.97      17.25   1.332 0.183383
scaled_dew_point                    107.82      30.38   3.550 0.000413 ***
scaled_wind_speed                   112.33      21.86   5.138 3.65e-07 ***
scaled_visibility                    79.96      17.39   4.599 5.08e-06 ***
scaled_sea_level_pressure           119.25      28.93   4.122 4.22e-05 ***
scaled_co                           -45.37      23.65  -1.918 0.055521 .
pm25_bin1                            89.66      45.56   1.968 0.049464 *
ozone_bin1                          184.35      50.40   3.657 0.000275 ***
scaled_no2                          -67.35      24.87  -2.708 0.006939 **
scaled_temp_aqi_interaction          66.48      36.86   1.804 0.071749 .
scaled_avg_temp_gender_interaction   55.64      88.36   0.630 0.529134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 401 on 672 degrees of freedom
Multiple R-squared:  0.9866,   Adjusted R-squared:  0.9862
F-statistic:  2478 on 20 and 672 DF,  p-value: < 2.2e-16
```

Table 7. Comparison of model performance on training data

Model	**RMSE**	**MAE**	**R^2**
Continuous PM2.5	394.664	295.4802	0.9866
PM2.5 Bin	394.8929	297.1224	0.9866

Table 7 shows a comparison of the models RMSE, MAE and R^2 on the training data which shows very similar model performance, with the model containing the continuous PM2.5 feature only slightly out-performing the model with the binned PM2.5.

Table 8 shows a comparison of the model performances on the test data. Unsurprisingly, the results are very similar. The model with PM2.5 left continuous has a negligibly lower RMSE and MAE, and a slightly higher R^2, which may make it a slightly preferable model for prediction.

Table 8. Comparison of model performance on test data

Model	**RMSE**	**MAE**	**R^2**
Continuous PM2.5	357.0011	267.278	0.9893897
PM2.5 Bin	358.0687	274.0892	0.9891762

Comparing Table 7 and Table 8, there is no obvious overfitting going on, as the model appears to perform better on the test data with a lower RMSE and MAE, and higher R^2. This will need to be reassessed as model adjustments and enhancements are put into place.

Discussion & Next Steps

Our preprocessing and feature-engineering steps gave us a good picture of how our data behaves and what we need to account for in our initial models. We imputed PM2.5 and removed PM10 due to missingness and created newly engineered features (*supershoe, temp_aqi_interaction, avg_temp_gender_interaction*). We then created bins for ozone and PM2.5 using decision-tree splits, revealing more stable patterns in how weather and

air quality relate to finishing times. These steps were driven by our question, how environmental factors predict average marathon finishing times across different performance groups and genders, and they helped us find which variables have the most interpretable effects.

The initial multiple linear regression models showed that marathon, gender, and subgroups still play an important role in explaining finish times, so including them as controls still makes sense for now. We ran two multiple linear regression models, with the only difference between models being the binned version of PM2.5 versus the scaled version of PM2.5. We found that most of the environmental and weather features were significant in predicting average finishing time, with a slowing effect. When the PM2.5 was continuous, we saw that it had a positive impact on finishing time, but when it was binned, it negatively impacted finishing times. Surprisingly, the model showed that increasing temperature led to faster race times, which may need to be investigated further. We did have two influential points when running the model assumptions, so we will have to look into those data points to see how they may be impacting the results, and if they are explainable. We may also consider investigating the models with unscaled data to see if that may be causing some of the counterintuitive results.

Next Steps: Model Selection & Tuning Plan

For the next stage of the project, we will move from simple linear regression models to models that are able to capture nonlinearities and interactions. We are going to use an XGBoost model, since tree-based models naturally capture nonlinearities and interactions without requiring us to create them. This could help us validate whether the interactions

we hand-selected align with the patterns the model finds. Hyperparameter tuning will focus on tree depth, learning rate, and the number of boosting rounds, while cross-validation will help ensure a good evaluation and help prevent overfitting.

Based on our current analysis, we may refine the feature set by including only the variables with clear signals to help reduce noise and improve interpretability through methods like forward, backward and stepwise selection. We also plan on testing a model with the original continuous features rather than the scaled features. This may also help us make adjustments to our analysis plan, specifically with how we handle interaction terms and whether additional engineered features are needed. Overall, by comparing multiple modeling approaches, including linear regression and XGBoost, we want to find the most accurate predictions and the most interpretable explanations for how environmental factors help predict marathon performance. We also plan to run a few models testing the model with and on the Berlin data as a case study.

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer.

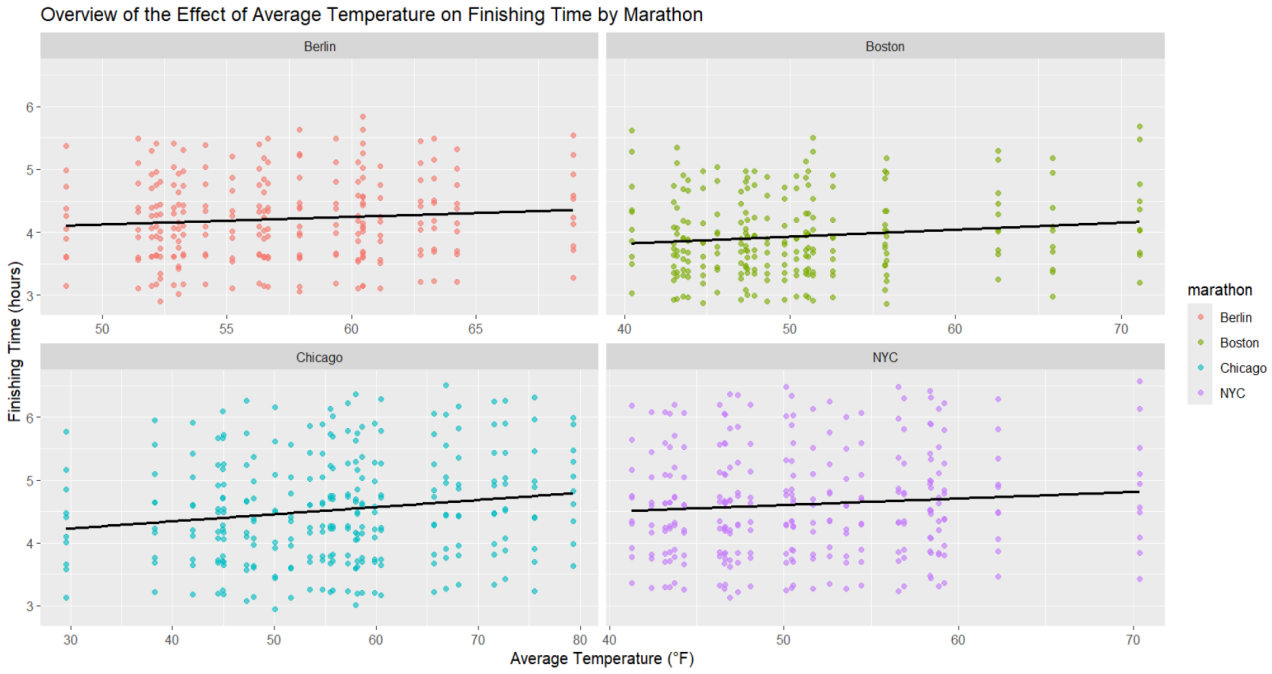
Appendix

Appendix 1. Data Dictionary

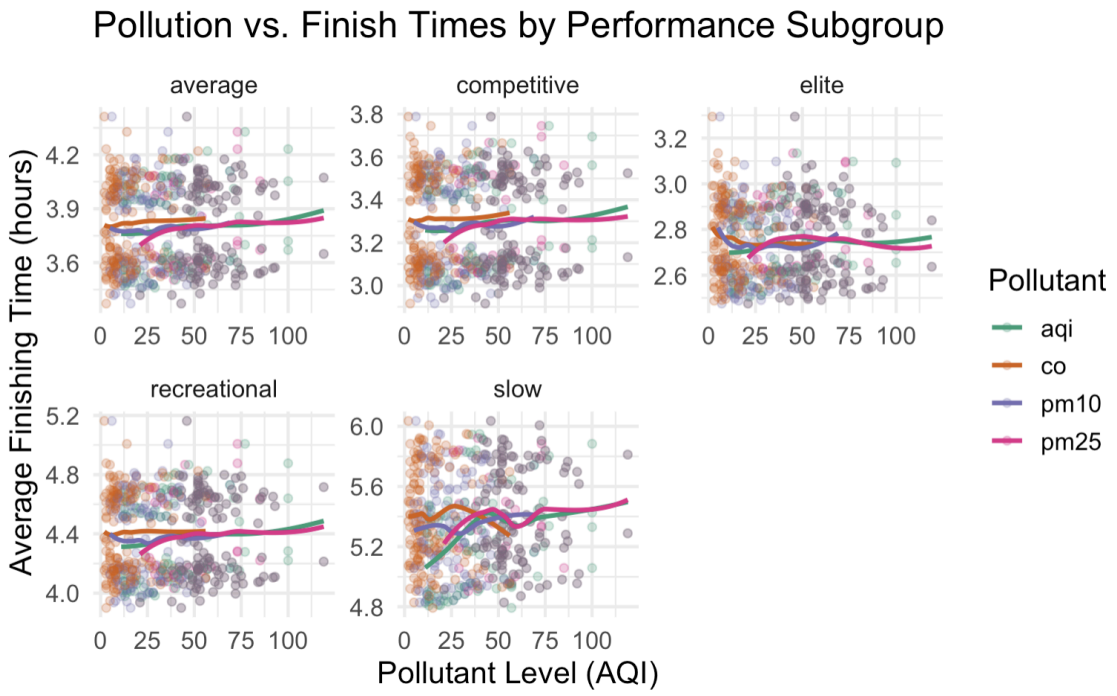
Variable	**Description**	**Type**
n	Total number of runners who completed the marathon	Integer
marathon	Name of the marathon event	Factor with 3 levels
year	Year the marathon occurred	Integer
gender	Gender category of runners (female/nonbinary and male)	Factor with 2 levels
subgroup	Running performance 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

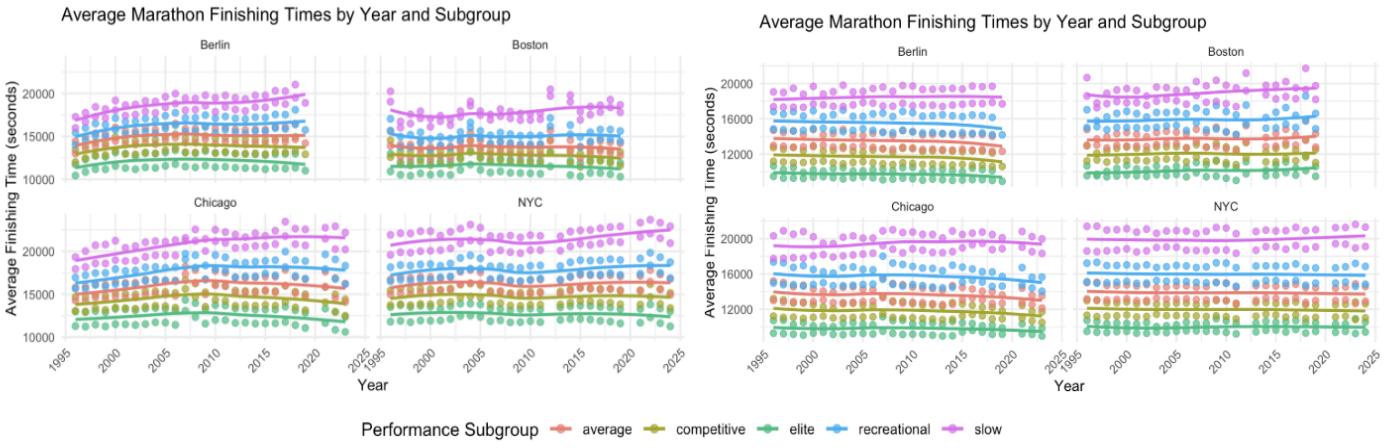
Appendix 2. Relationship between average temperature and finishing time by each marathon.



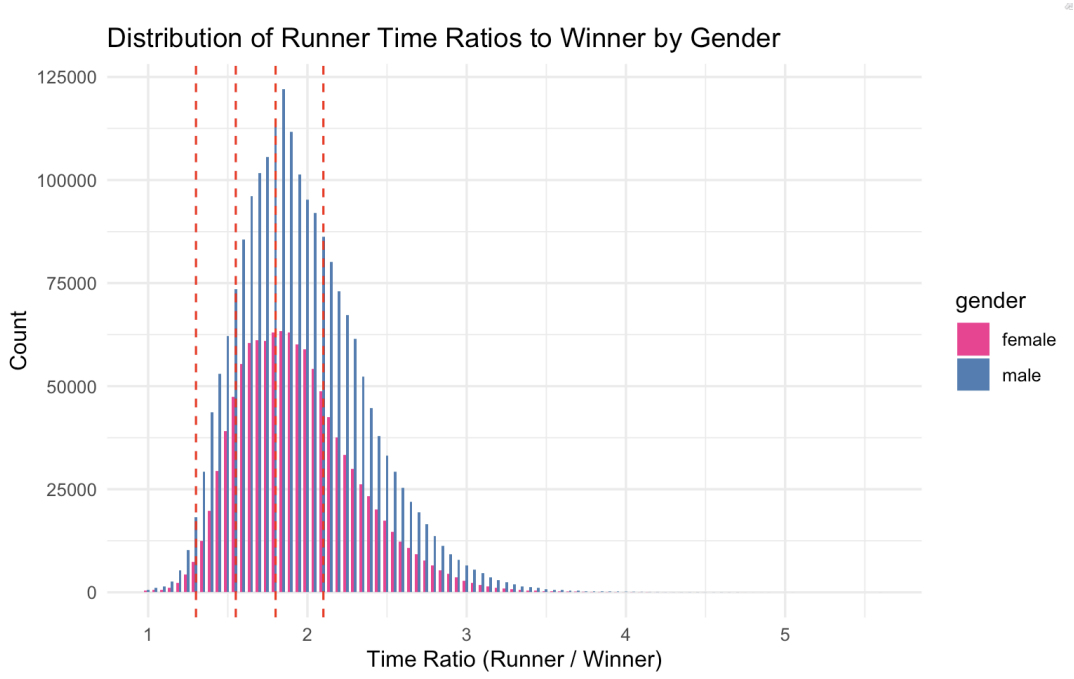
Appendix 3: Scatterplot showing the relationship between pollutant levels and finishing times by subgroups.



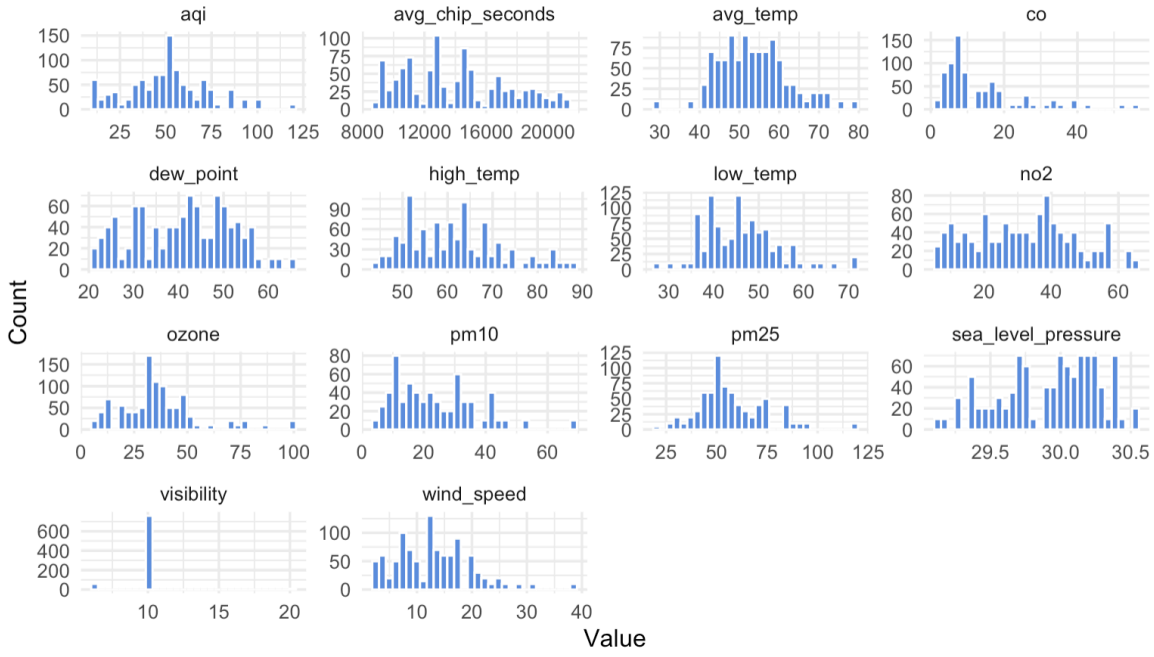
Appendix 4: Scatter plots by marathon and subgroup. The scatter plot on the left represents the quintiles, and the one on the right represents the new performance-based cutoffs. Colors represent the different performance groups.



Appendix 5: 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).



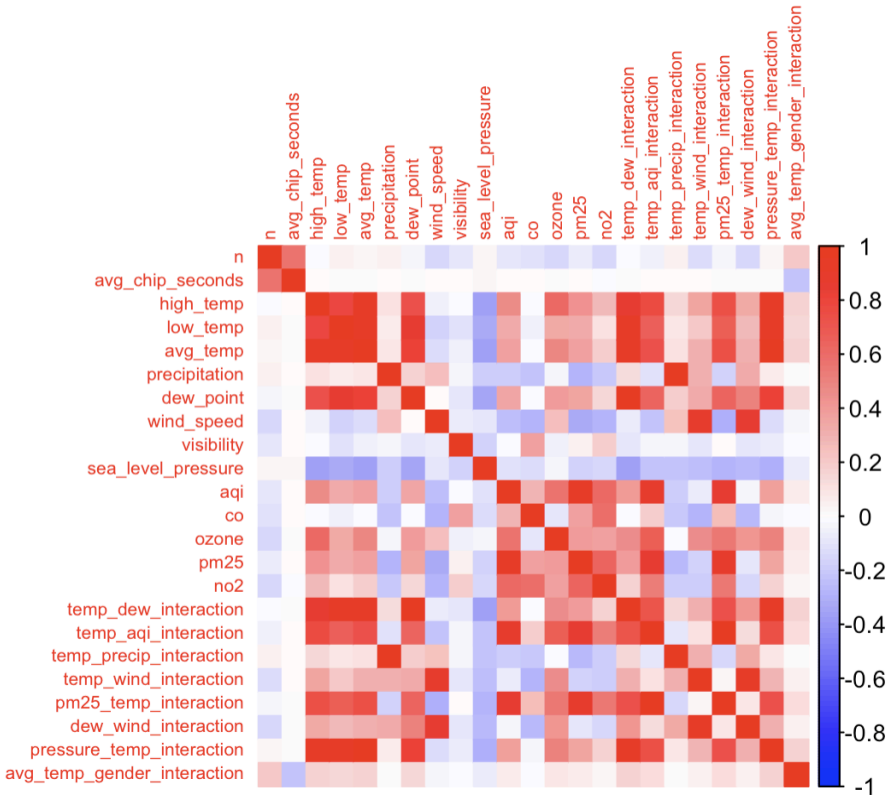
Appendix 6: Histograms of Continuous Variables



Appendix 7: 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 8. Correlation matrix that includes the newly engineered features.



Appendix 9: Table of 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

Appendix 10. 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 11. 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
```