Preprocessing & Feature Engineering Report

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

Team Lead: Krisha Bugajski

Recorder: Zachary D'Urso

Spokesperson: Meghan Holden

Link to GitHub:

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

Table of Contents

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).

Methods

Restatement of previous plan

Last week, our plan for preprocessing and feature engineering included standardizing marathon datasets, handling missing values, normalizing continuous variables, and creating interaction terms and control variables such as supershoes and COVID-19 disruptions. After completing exploratory data analysis, results showed that pollution and temperature decreased the performance of runners across all subgroups. Results also suggested that the method of creating performance subgroups needed to be reworked. This is important, as part of our feature engineering plan included finalizing control variables for both the introduction of "super shoes" beginning around 2018 and the onset of COVID-19-related race disruptions starting in 2020. However, when we plotted

subgroups over time, we did not see any speed up related to supershoes at all, and instead saw average finishing times increasing as years progressed, suggesting that the quintiles method of subgrouping may not be appropriate. Because of this, we plan to define performance groups based on ratios of each runner's time to the winner's time, which allows consistent subgroup comparisons across years, genders, and marathons. We also observed substantial missing environmental data for Berlin, leading us to treat Berlin as a separate case study rather than including it in the main modeling dataset.

Preprocessing Methods

We started our preprocessing steps by merging the marathon, weather, and air quality datasets into a single dataset, ensuring that variables had consistent names and formats. We inspected variable types to confirm that continuous predictors were stored as numeric and categorical variables such as gender, marathon, and subgroup were all set as factors so that the models could interpret them correctly. Because Berlin contained substantial missing environmental measurements (CO, PM10, PM2.5, and visibility), we chose to separate Berlin from the main modeling dataset and evaluate it later as a secondary case study. For the rest of the marathons, we plan to see the data missingness patterns and decide whether it will be best to perform imputations or the removal of variables.

We will also look at the histograms for all the continuous variables to see distribution shapes and confirm that finishing time conversions were correct. Since the models we plan on using, multiple linear regression and gradient boosting, handle scaled numeric features effectively, we will scale all numeric predictors except the outcome variable (avg_chip_seconds) and identifiers such as year. Before modeling, we will finalize a

training test split of approximately 90–10%, since the removal of Berlin reduces our sample size, and a larger training proportion will help maintain model stability while still allowing us to evaluate predictive performance.

Feature Engineering Methods

Originally, our feature engineering plan included creating interaction terms such as temperature × dew point and AQI × temperature, adding a supershoe and COVID-19 control variables coded by year, and generating consistent subgroup categories to compare performance across runner types. However, EDA revealed that the quintile-based subgrouping produced misleading temporal trends due to increasing participation among slower runners. Therefore, we revised the subgroup definitions to use runner-to-winner time ratios, a method that is more stable across years. This change helps avoid the distortion that is caused by the growing number of slower participants and helps keep subgroups comparable across marathons.

We also planned to reduce multicollinearity by examining correlations and removing redundant predictors rather than applying PCA, since our features have clear environmental meaning. From EDA, we determined that AQI is highly correlated with PM2.5, so AQI will be removed. PM10 will also be excluded because this variable contains a lot of missing data. After removal, we will finalize engineered features and scale the remaining numeric predictors.

Results

Subgroup Definition

Runner performance was divided into five subgroups (elite, competitive, average, recreational, slow) based on the ratio of individual finishing time to the winner's time within each marathon year gender group. Average finishing times per subgroup were then computed for analysis. Figure 2 illustrates how quintile-based groups showed an apparent increase in average finishing times over the years, whereas the performance-based cutoffs eliminate this trend.

It is important to note that the resulting distribution is skewed, where faster groups (elite, competitive) have fewer runners, and the slower groups dominate the counts, as seen in Figure 1. Based on the distribution, subgroup cutoffs were selected to fit with natural patterns in the data. The histogram reveals clear clustering in runner-to-winner time ratios and a long right-skewed tail, which helped with the placement of the performance thresholds (1.30, 1.55, 1.80, 2.10). So these cutoffs are not arbitrary, and they follow the natural shape of the data rather than relying on fixed percentile breaks. For the purposes of modeling, we left the subgroup sizes as they were. This imbalance is noted so readers understand why the “slow” trends dominate counts. Weighting could later be applied; however, it is not initially because subgroup effects are not the primary focus. Female counts are lower overall, reflecting participation differences, but both genders show similar distributional shapes. Overall, this figure supports the use of ratio-based performance categories by illustrating natural clustering and skewness in marathon finishing times.

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).

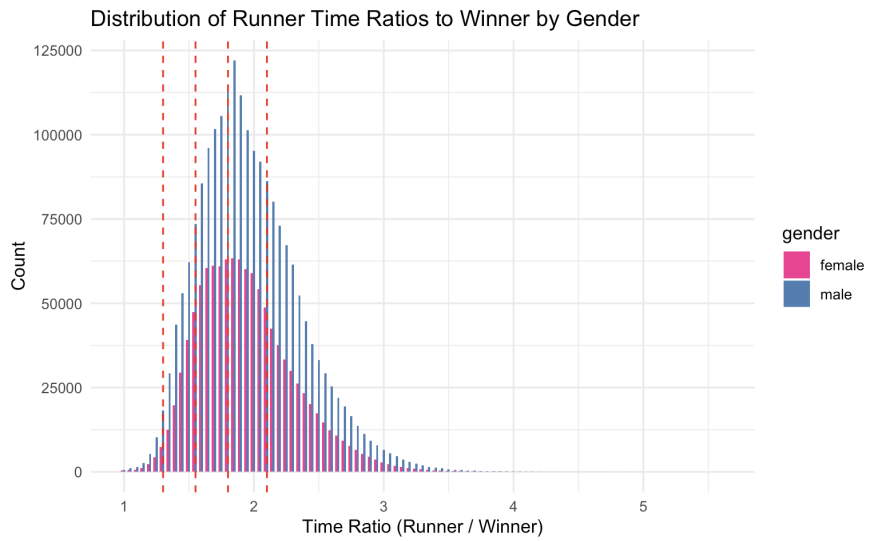
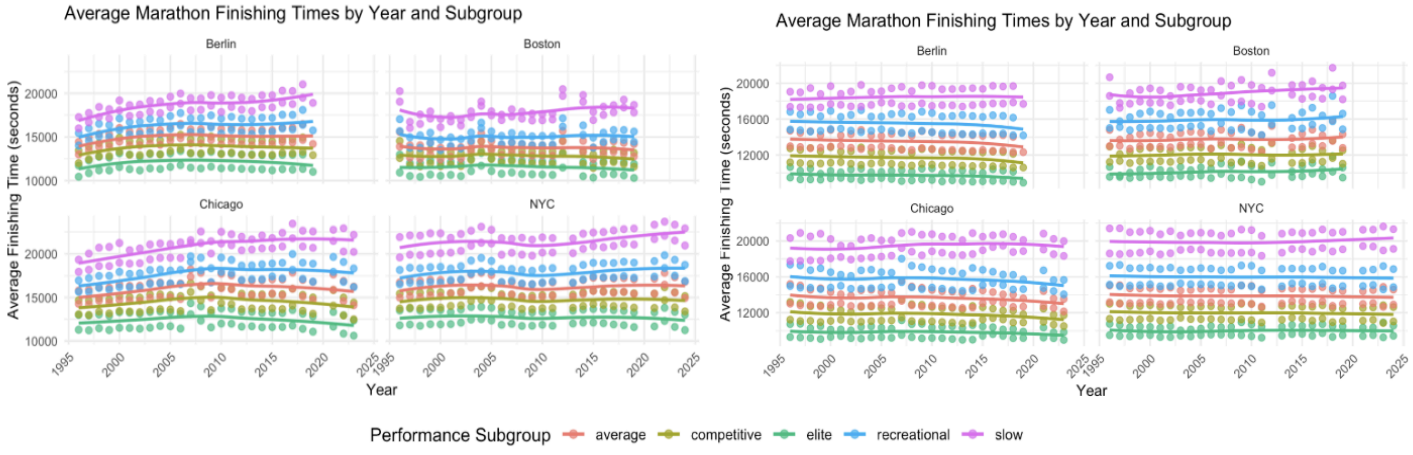


Figure 2: Scatter plots by marathon and subgroup. The scatter plot on the left represents the quintiles, and the one on the right represents the new performance-based cutoffs. Colors represent the different performance groups.



Missing data

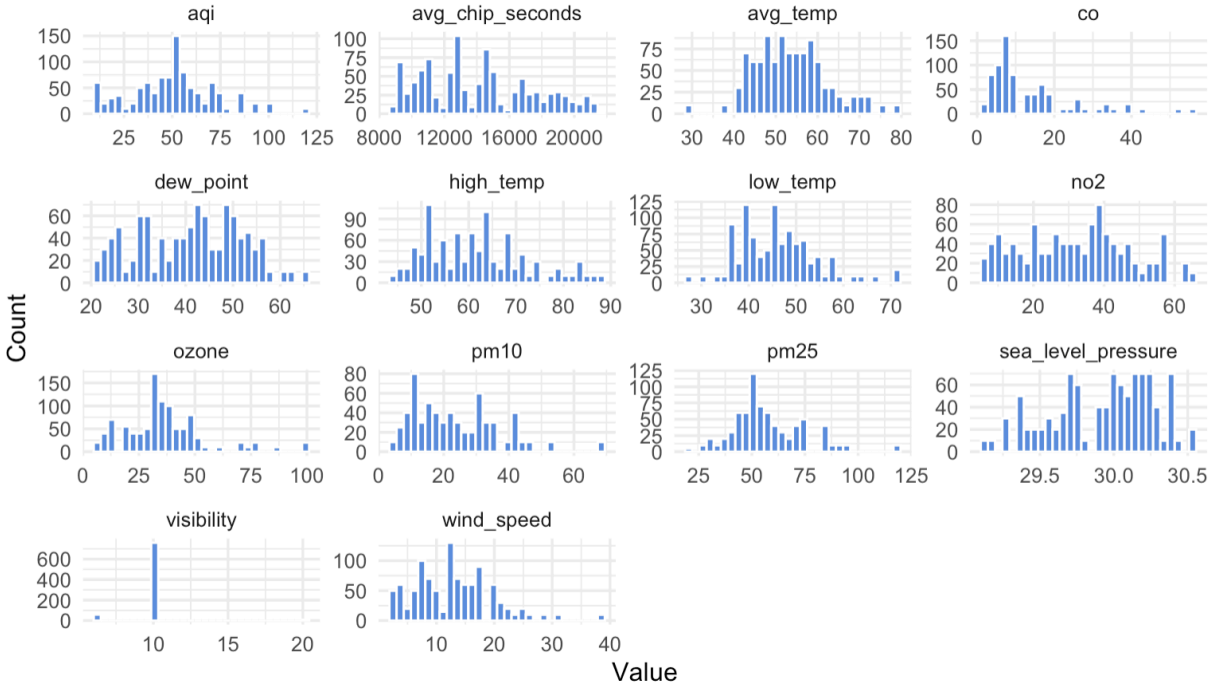
After separating Berlin from the other marathons, we checked how much information was missing in our main dataset (770 observations, 21 variables). We found that PM10 had a large amount of missing values (320), so keeping it would have added more noise than value, meaning it made sense to drop that variable altogether.

PM2.5, however, is an important pollutant and had fewer missing values (70), so instead of removing it, we filled the gaps using K-nearest neighbors (KNN) imputation. We choose to use KNN because it works by finding rows that look similar and using them to estimate the missing value, which lets us keep the natural structure and variability in the data. Since our dataset is fairly small, this method works well and avoids overcomplicating things. We started with $k = 5$, but we plan to test other k values later to see which one gives the best model performance.

Histograms of Continuous Variables

We looked at histograms for all the continuous variables to see distribution shapes and confirm that finishing time conversions were correct, as seen in Figure 3. While several environmental variables did display expected right skew (e.g. PM2.5, wind speed), no transformation is currently planned for our modeling goals. Finishing times were also right-skewed, which is typical for marathon data.

Figure 3. Histograms of Continuous Variables



Feature engineering

We created several interaction terms to try and capture potential combined effects of environmental and demographic factors on marathon performance, as summarized in Table 1, knowing that we might drop them due to high correlations. 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. In addition, we added a control for the introduction of "supershoes," coded as 1 for marathons in 2018 and later, when these performance-enhancing shoes became widely used, and 0 otherwise. Including this variable allows the models to account for technological improvements in footwear that could affect finishing times across years.

Table 1: 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)

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 4. We quickly noticed that a lot of these features were highly correlated, as seen in Table 2. Therefore, we decided to remove all the features that introduced multicollinearity, including ones that we found to be highly correlated during our initial EDA.

Amongst the newly engineered numerical features, we decided to keep **temp_aqi_interaction.** The *temp_aqi_interaction* combines temperature and aqi, allowing the model to account for their joint effects while avoiding redundancy with other variables. Although this interaction has a moderate correlation of 0.74 with *avg_temp*, we are keeping it because it captures meaningful combined effects that are not fully represented by temperature alone, with its preservation of *aqi*.

Amongst the categorical features, we decided to keep **supershoe** and **avg_temp_gender_interaction**. *Supershoe* captures the post-2018 period, providing an important categorical indicator for potential temporal shifts. The *avg_temp_gender_interaction* models how the effect of temperature may differ by gender, capturing potential variations in the outcome across subgroups.

Figure 4. Correlation matrix that includes the newly engineered features.

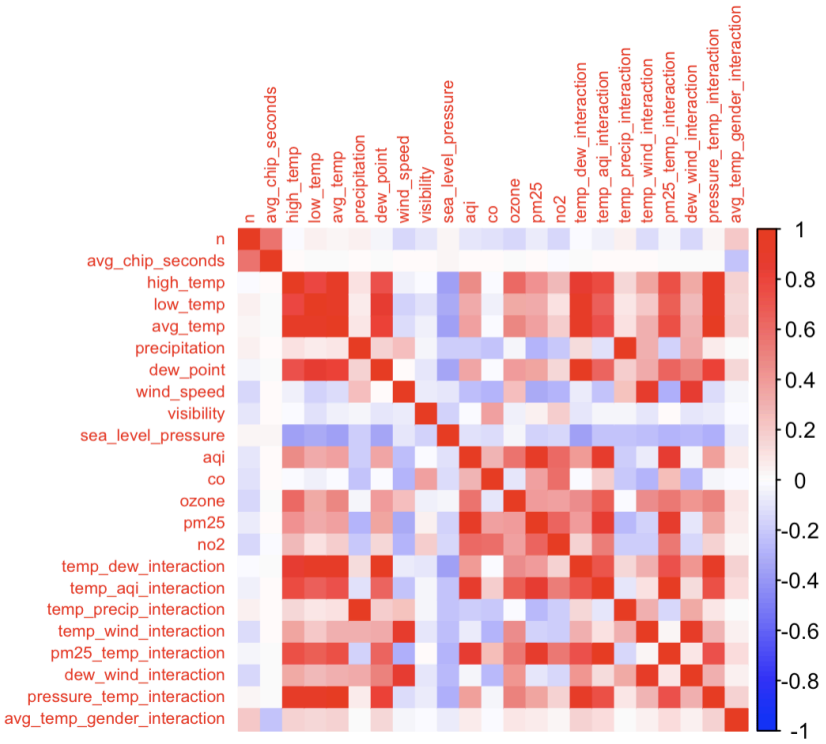


Table 2: Features to drop.

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

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 Plan

During the initial modeling phase, we will explore additional feature engineering using tree-based methods. This will help determine whether certain continuous variables are better left as continuous or grouped into categories based on meaningful thresholds. Overall, this approach will help capture nonlinear relationships and interactions. We will additionally try regularization methods such as LASSO and tree-based variable importance to help further with feature selection.

Feature Scaling

We then scaled all numeric predictor variables (except the outcome, avg_chip_seconds, and identifiers, n and year) so they have a mean of 0 and a standard deviation of 1, as seen in Appendix 2. This puts predictors on the same scale, making model coefficients more interpretable and helping the algorithms be more reliable. Both the original and scaled versions were kept in the dataset for flexibility in modeling.

Final Pre-Processing Steps

One of the last steps before heading into the initial model is to break the data into test and training data. The test data set will consist of 10% of the data held out, leaving 90% of the data for the model to be trained on. Our original plan was an 80-20 training-test split; however, due to the Berlin data being investigated as a case study, which may not be part of our final data for model training, we decided to switch to 90-10 so that more data is available for training.

Discussion & Next Steps

Discussion

After processing the data and developing new features, the data is ready to train an initial model to predict average finishing times across different performance groups based on environmental factors and weather. Our primary research question is how environmental factors influence average finishing times across runner performance groups, and the following discussion summarizes how preprocessing and feature engineering prepare the data to answer this question.

Adjustments were made from the EDA, beginning with the way that subgroups were developed. The updates to the group, based on the winner-to-runner ratio, show the anticipated improvement around the introduction of super-shoes, with the exception of slow runners who do not see any improvement, as shown in Figure 2. The development of interaction terms and reassessment of correlations led to the removal of some features initially found in the dataset due to high correlation, reducing noise and multicollinearity (e.g. high and low temperatures), which also ensures that model coefficients remain

interpretable. While the interaction term temp_precip_interaction was initially created, it was quickly removed after reviewing the correlation matrix due to high correlation (as seen in Table 2).

PM10 and PM2.5 had higher numbers of missing values coming out of EDA. Because PM2.5 is an important pollutant and has fewer missing values than PM10, the decision was made to remove PM10 and impute PM2.5 using K-nearest neighbors (KNN) imputation. Finally, numeric features were scaled to allow the model to be more widely applicable. Further feature selection and potential transformations will be guided by tree-based models and regularization methods, such as LASSO, during the modeling phase.

Modeling Plan

Our modeling plan is to compare multiple linear regression models with tree-based gradient boosting approaches, such as XGBoost or LightGBM, to determine which best captures the relationships relevant to our research question. Initial tree-based models will also investigate whether additional feature engineering of continuous variables would benefit the model overall, including potential threshold-based grouping. Regularization methods such as LASSO and tree-based variable importance will also be performed to help with final feature selection before heading into our final modeling.

The model will be evaluated using the training and test sets to evaluate model performance, using metrics such as R^2, MAE, and RMSE using cross-validation. We will then visualize actual vs. predicted average finisher times, review feature importance, and check residuals for bias. Finally, we will interpret how factors like temperature, wind,

dew point, and air quality influence average finish times among each of the performance groups, using these results to refine the model and make sure the model's results are easy to understand.

Appendix

Appendix 1. Data Dictionary

Variable	Description	Type
n	Total number of runners that completed the marathon	Integer
marathon	Name of the Marathon Event	Categorical
year	Year of the marathon	Integer
gender	Gender category of runners	Categorical
subgroup	Running performance category: elite, competitive, average, recreational, slow	Factor w/ 5 levels
avg_chip_seconds	Average finishing chip time in seconds by marathon, subgroup, and gender	Numerical
high_temp	Daily high of race day (°F)	Integer
low_temp	Daily low of race day (°F)	Integer
avg_temp	Daily average temperature (°F)	Numerical
precipitation	Total precipitation on race day (in)	Numerical
dew_point	Average dew point temperature (°F) on race day	Numerical
wind_speed	Average wind speed on race day (mph)	Integer
visibility	Average visibility on race day (statute miles)	Numerical
sea_level_pressure	Average sea level pressure on race day (inHg)	Numerical
aqi	Average Air Quality Index (AQI) on race day	Integer
main_pollutant	Primary pollutant on race day	Categorical
co	Average carbon monoxide (AQI) on race day	Numerical
ozone	Average ozone (AQI) on race day	Numerical
pm10	Average particulate matter with diameter of less than 10 micrometres (PM10) (AQI) on race day	Numerical
pm25	Average particulate matter with diameter of less than 2.5 micrometres (PM2.5) (AQI) on race day	Numerical
no2	Average nitrogen dioxide (AQI) on race day	Numerical

Appendix 2. Numeric scaled variables

```
 scaled_avg_temp   scaled_precipitation scaled_dew_point   scaled_wind_speed  scaled_visibility
 Min.   :-2.4778   Min.   :-0.2349      Min.   :-1.60752   Min.   :-1.57333   Min.   :-0.1146
 1st Qu.:-0.6844   1st Qu.:-0.2349      1st Qu.:-0.75144   1st Qu.:-0.56935   1st Qu.:-0.1146
 Median :-0.1846   Median :-0.2349      Median :-0.02239   Median :-0.06737   Median :-0.1146
 Mean   : 0.0000   Mean   : 0.0000      Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
 3rd Qu.: 0.5887   3rd Qu.:-0.2349      3rd Qu.: 0.67812   3rd Qu.: 0.43462   3rd Qu.:-0.1146
 Max.   : 2.8604   Max.   : 6.2934      Max.   : 2.45103   Max.   : 3.94854   Max.   : 8.7121
 scaled_sea_level_pressure   scaled_co        scaled_ozone      scaled_pm25        scaled_no2
 Min.   :-2.10697          Min.   :-1.0143   Min.   :-1.3289   Min.   :-1.8666   Min.   :-2.42141
 1st Qu.:-0.68559          1st Qu.:-0.6748   1st Qu.:-0.7027   1st Qu.:-0.5950   1st Qu.:-0.70035
 Median : 0.08196          Median :-0.4201   Median :-0.3269   Median :-0.2317   Median : 0.03725
 Mean   : 0.00000          Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 0.90636          3rd Qu.: 0.2591   3rd Qu.: 0.2993   3rd Qu.: 0.4950   3rd Qu.: 0.61094
 Max.   : 1.84447          Max.   : 3.5700   Max.   : 3.6183   Max.   : 3.7649   Max.   : 2.33200
 scaled_temp_aqi_interaction scaled_avg_temp_gender_interaction
 Min.   :-1.2468             Min.   :-0.9695
 1st Qu.:-0.6611             1st Qu.:-0.9695
 Median :-0.2677             Median :-0.4251
 Mean   : 0.0000             Mean   : 0.0000
 3rd Qu.: 0.1970             3rd Qu.: 0.9061
 Max.   : 3.0380             Max.   : 1.9510
```