Project 1: Exploratory Data Analysis

Jiahao Cui

Abstract

This study examines how environmental factors, including temperature, humidity, solar radiation, wind, and WBGT, air quality impact marathon performance across ages and genders. We introduced the adjusted percentage current course record to normalize finish times for a more accurate comparison. Performance generally improves from ages 15 to 30, followed by a steeper decline in females than in males. Environmental conditions significantly influence performance, with stability in good weather but a more pronounced decline with age in bad weather. Older individuals are especially sensitive to adverse conditions, and solar radiation and air quality were identified as the most influential factors, showing a complex, non-linear relationship with performance.

Introduction

The purpose of this study is to investigate how environmental factors, temperature, humidity, solar radiation, and wind, impact marathon performance across the lifespan in both men and women, with the hypothesis that older individuals will experience more pronounced slowing in high WBGT conditions compared to younger individuals, and similar trends will be observed in both men and women.

Previous research has shown that endurance performance declines as environmental temperatures rise[1], with this decline being more pronounced in longer-distance events like marathons (42.2 km)[2]. Additionally, older adults face greater thermoregulatory challenges, limiting their ability to effectively manage heat, which can worsen the effects of higher temperatures[3]. Furthermore, sex differences in both endurance performance[4] and physiological responses to heat are well-established[5].

The remainder of this report is structured as follows: Data Collection, Data Preprocessing, answering three aims of the study, and References.

Table 1: Summary table of environmental conditions

```
# A tibble: 5 x 11
   Race Avg_Age Avg_CR Avg_Td Avg_Tw Avg_rh Avg_Tg Avg_SR Avg_DP Avg_Wind
  <int>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>    <dbl>
1     0    47.0   41.2   11.6   7.59   36.1   24.2   650.   3.32    12.0
2     1    45.8   51.6   12.4   8.54   60.4   24.5   460.   4.65     8.21
3     2    49.6   54.9   11.7   7.57   26.9   21.4   401.   2.74    11.2
4     3    44.7   45.7   13.1   9.85   41.8   24.9   436.   5.96     8.80
5     4    44.0   47.6   18.9  14.9    49.3   31.6   677.  12.4      9.16
# i 1 more variable: Avg_WBGT <dbl>
```

Data Collection

The primary dataset, *project1.csv*, includes top single-age performances from men and women, aged 14 to 85, across five major marathons: the Boston Marathon, Chicago Marathon, New York City Marathon, Twin Cities Marathon (Minneapolis, MN), and Grandma's Marathon (Duluth, MN), spanning 15 to 20 years (1993–2016). The dataset also provides detailed environmental conditions for each event.

The environmental conditions are represented by various variables, including temperature, humidity, solar radiation, and wind. These are captured as: "Td..C" (Dry Bulb Temperature in Celsius), "Tw..C" (Wet Bulb Temperature in Celsius), "X.rh" (Percent Relative Humidity), "Tg..C" (Black Globe Temperature in Celsius), "SR.W.m2" (Solar Radiation in Watts per square meter), "DP" (Dew Point in Celsius), "Wind" (Wind Speed in km/h), and "WBGT" (Wet Bulb Globe Temperature). The first seven variables represent key environmental indices. "WBGT" is calculated as a linear combination of temperature variables:

$WBGT = (0.7 * Tw..C) + (0.2 * Tg..C) + (0.1 * Td..C)$

Additionally, "Flag" is a categorical index based on WBGT values: White for WBGT < 10°C, Green for WBGT between 10-18°C, Yellow for WBGT > 18-23°C, Red for WBGT > 23-28°C, and Black for WBGT > 28°C. The summary of these environmental condition variables is presented in Table 1.

Additional datasets used in this project is *aqi_values.csv* which contains the air quality index (AQI) for each marathon, sourced from the United States Environmental Protection Agency[8].

Data Preprocessing

Firstly, from Table 1, we observe that there are 491 missing values in certain variables, which are primarily from races in Chicago, New York City, Twin Cities (2011), and Grandma's Marathon (2012). Additionally, the maximum WBGT value is 25°C, meaning no black flag (WBGT $>$ 28°C) is present in the data. Due to limited data for the youngest and oldest runners, we will filter the dataset by age, focusing on participants aged 20 to 80 in certain scenarios.

Secondly, the variable "X.CR" represents the percentage off the current course record by gender, which serves as a measure of marathon performance. Therefore, we need to integrate the current course record data from course_record.csv to calculate the actual marathon finishing times.

We also want to point out directly use "X.CR" will cause bias, because "X.CR" is make a normalization on the finishing time and current course record. However, the course record is a non-increasing value over time, meaning the best finish time in each year is a form of missing data that is only observed when someone breaks the record. Using the normalized "X.CR" as a representation of marathon performance each year can lead to bias. For instance, consider a race in Boston in 2000 and 2015, where the weather conditions and the best finish times in a specific age and gender group are identical. In 2000, the record was 2.1 hours, while in 2015 it was 2 hours. Even though the finish times are the same, the "X.CR" values differ significantly. If we could observe the best marathon finish time each year, it would serve as a better estimator of performance under the given conditions.

To avoid this bias, one viable solution is to calculate the finish time instead of using "X.CR." However, finish times vary across different race locations, not only due to weather, but also because of the terrain and course design vary. Separating the data by marathon location might reduce robustness, and since our research does not focus on location differences, considering all data together is preferable.

Thus, we propose a new performance index: the Adjusted Percentage Current Course Record (Adjust_X.CR). This index uses the average course records over the years for each race and gender to represent performance independently of environmental conditions. We calculate Adjust_X.CR using the formula:

$Adjust\_X.CR = X.CR * (CR + 1)/Mean\_Record$

This adjusted value ensures consistency, meaning the same Adjust_X.CR corresponds to the same finish time across different years, allowing for performance comparisons across locations.

Finally, we merge the air quality index (AQI) data with the updated dataset. The "mean_aqi" is the average AQI for the "8-HR RUN AVG BEGIN HOUR," which is another weather-related index. Since AQI data is missing for "1-Hours run", we filter the data by the "8-HR RUN AVG BEGIN HOUR" to obtain the relevant information.

Table **??** illustrates the differences between "X.CR" and "Adjust_X.CR". Additionally, we present the statistics for "mean_aqi", which serves as a weather condition index.

AIM 1:

Examine effects of increasing age on marathon performance in men and women

In this case, we ignore the different in weather condition, and only focus on the effect of age and sex on marathon performance. We will calculate the mean, max, and min(through different races and years) of Adjusted Percentage Current Course Record (Adjust_X.CR) for each age and sex group.

Figure 1 shows that marathon performance increases from ages 20 to 30, followed by a decline in both female and male groups. The decline is more pronounced in females compared to males. The larger error bars in the older age groups may be attributed to the limited data available for these ages. This trend is similar to the plot presented in the slides, but in this case, we used Adjust_X.CR instead of X.CR.

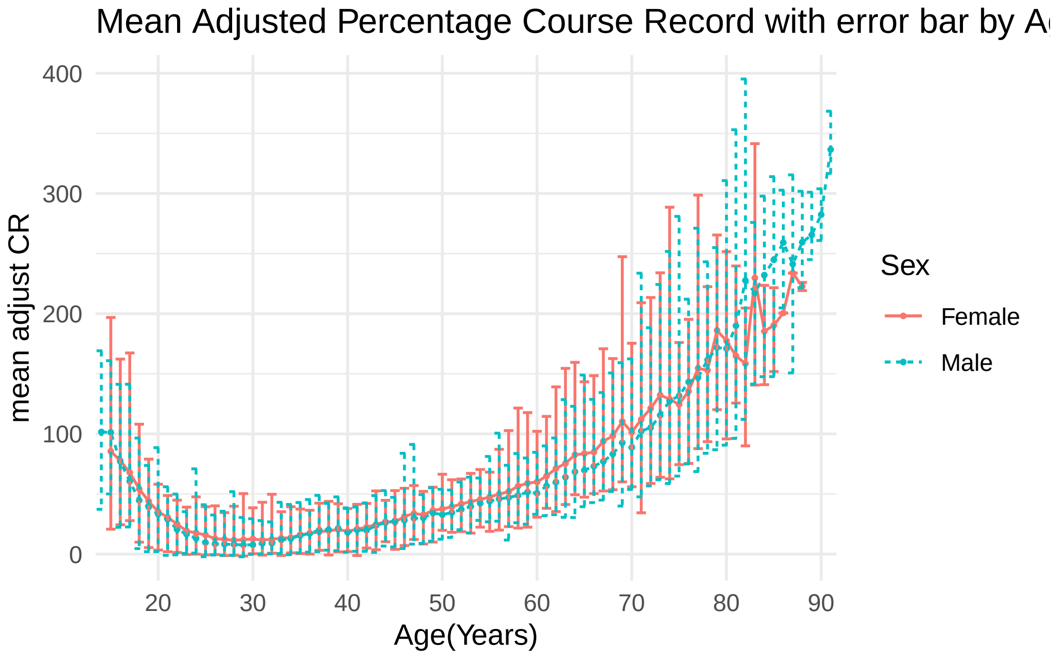


Figure 1

We also fit the mean values using a quadratic polynomial to illustrate the trend in Figure 2. This shows that the trend is not linear, and the best-fit lines do not accurately reflect the ages of peak performance.

We excluded data for ages below 20 and above 70 due to limited representation in these groups. Based on this filtered data, the best performance is observed between ages 25 to 33 for males and 28 to 32 for females. However, the peak of the fitted curves occurs at age 31 for males and 32 for females. Despite this, the differences between the actual data and the fitted curves are much smaller compared to the plot in the slides.

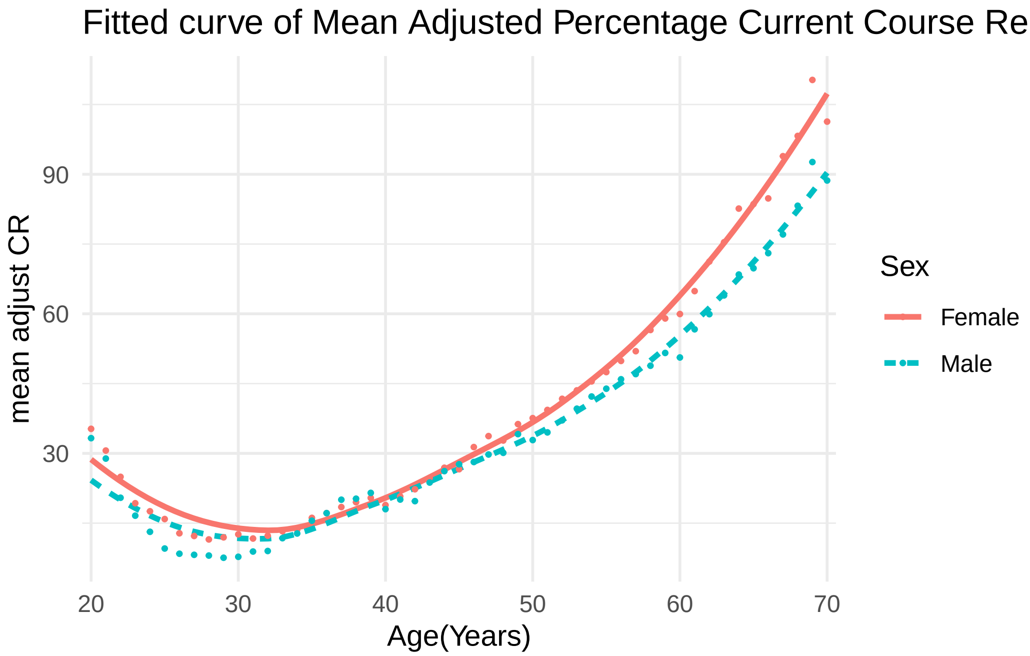


Figure 2

Conclusion:

Marathon performance increases from ages 15 to 30, followed by a decline. The decline is more significant in females compared to males. Larger error bars in the older age group are likely due to the limited data available for those ages. The best-fit lines do not accurately reflect the ages of peak performance, likely due to the limited data in the older groups and the fact that the trend is neither strictly linear nor quadratic. Additionally, using Adjust_X.CR instead of X.CR improves the overall fit.

AIM 2:

Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

The environmental conditions consist of eight variables. WBGT is a linear combination of three of these variables, and Flag is a categorical bin derived from WBGT. Instead of relying solely on Flag, we will cluster individuals using all the covariates to avoid losing valuable information.

We will first perform dimensional reduction to facilitate data visualization (EDA). After normalizing the data, we use K-means clustering[6] and the optimal number of clusters is determined by the elbow method, as shown in Figure 9, which identifies five as the optimal number of clusters.

Table 4 illustrates the relationship between Cluster and Flag, showing that clusters 3, 4, 6, and 7 correspond closely to Flags White, Green, Yellow, and Red, respectively. We then plot the clusters with Flag labels in Figure 3. Based on this, we can define cluster 6 as representing good weather, cluster 4 as representing bad weather, and clusters 3 and 7 as representing normal weather.

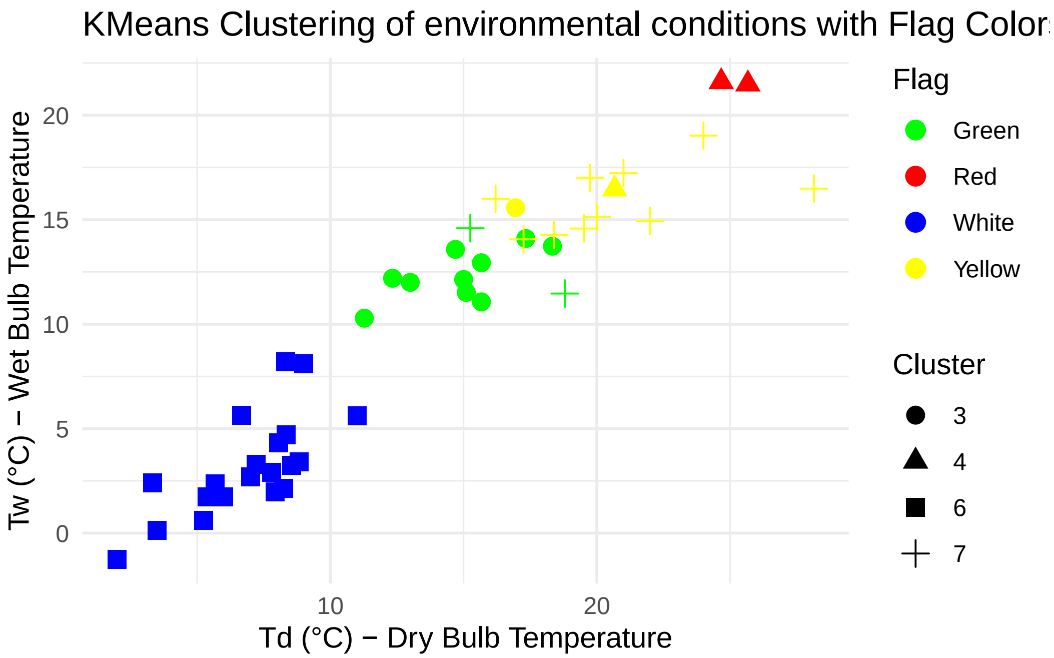


Figure 3

We can plot Mean Adjusted Percentage Course Record against age and gender, grouped by different weather conditions. Figure 4 shows that marathon performance is better in good weather (cluster 6) and worse in bad weather (cluster 4). However, the difference between good and normal weather is not significant.

We can also calculate the performance difference between ages, as shown in Figure 5, using the following formula:

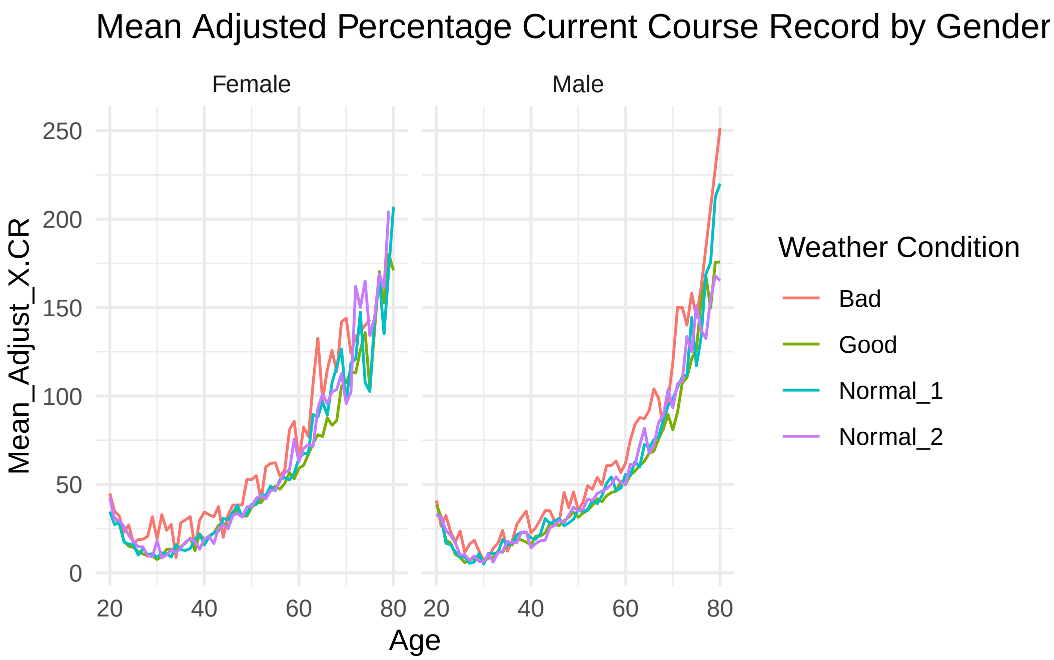


Figure 4

$$Difference\ Adjust\ X.CR = |Adjust\ X.CR(Age+1) - Adjust\ X.CR(Age)|$$

The results show that in bad weather (cluster 4), the performance difference between ages is more pronounced compared to other clusters. In contrast, good weather conditions (cluster 6) exhibit the smallest performance difference between ages. This indicates that weather conditions significantly impact marathon performance, and this impact varies across age groups.

Conclusion:

The analysis reveals that environmental conditions significantly affect marathon performance, with weather playing a crucial role in the variation across different age groups. In good weather (cluster 6), performance remains relatively stable across ages, whereas in bad weather (cluster 4), the decline in performance with age is more pronounced. The most significant performance differences are observed in extreme weather conditions, indicating that older individuals are more sensitive to adverse weather effects.

AIM3:

Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

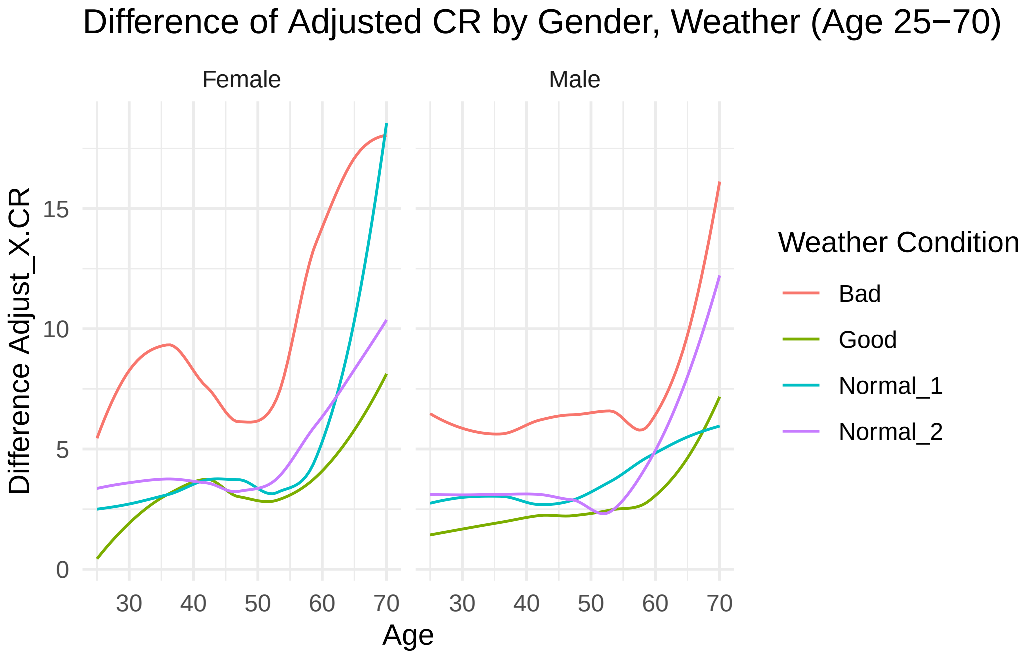


Figure 5

We include the air quality index (AQI) data in this part, as AQI is also a relevant weather condition index. We calculate the correlation matrix between performance metrics and various weather variables. The correlation between the Mean Adjusted Percentage Current Course Record and age, along with nine weather variables, is presented in Table 2.

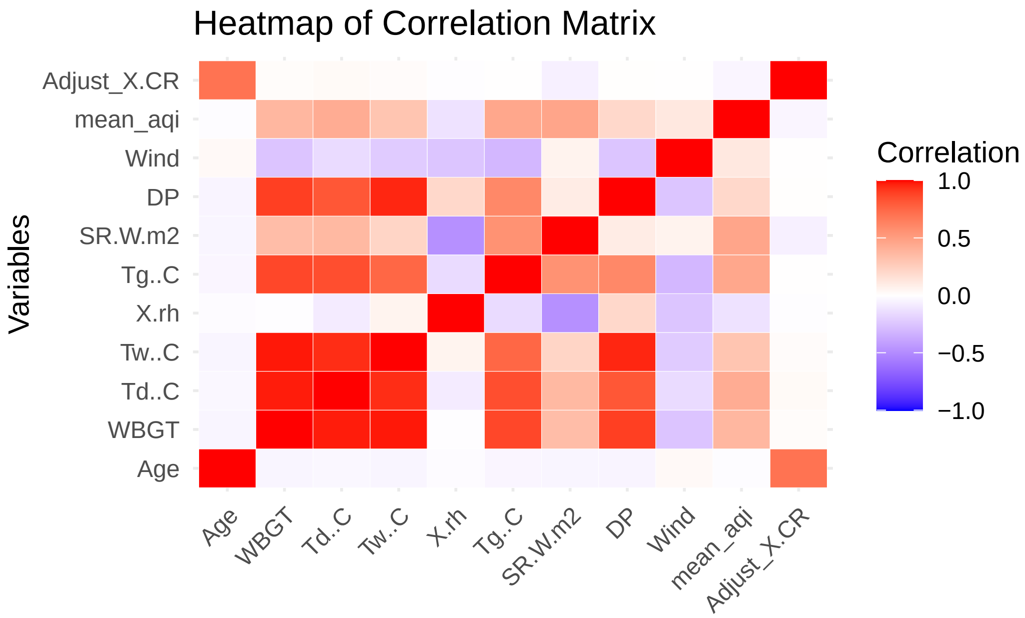
The results show that marathon performance is strongly positively correlated with age, consistent with our previous analysis. Therefore, we calculate the correlation between weather conditions and performance, adjusting for age.

Figure 6 illustrates the correlation between the weather variables and Mean Adjusted Percentage Current Course Record across different age groups. This helps identify which variables have the strongest linear relationships with performance.

Four weather variables show a strong positive correlation with Mean Adjusted Percentage Current Course Record: Wet Bulb Temperature (Celsius), Dry Bulb Temperature (Celsius), Black Globe Temperature (Celsius), and Dew Point (Celsius). The remaining four variables do not exhibit a strong correlation with Adjust_X.CR, although there is a strong negative correlation with Adjust_X.CR as age increases.

Next, we use statistical models to identify the most significant weather parameters affecting marathon performance. We begin with a linear regression model, where the performance metric is the dependent variable, and weather parameters (such as WBGT, temperature, and Flag) are the independent variables. Interaction terms between age, sex, and weather variables

Table 2: Table of correlation matrix



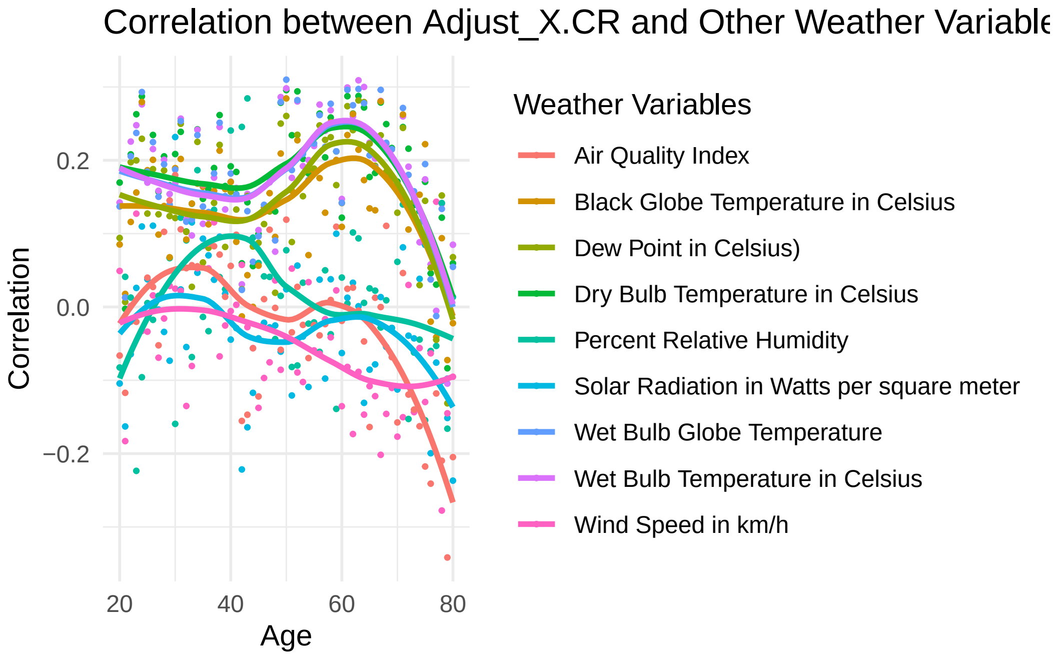


Figure 6

Table 3: Table of random forest, importance of variables(larger value means more important)

```
                    Age   Gender    Td..C    Tw..C     X.rh    Tg..C  SR.W.m2
IncNodePurity 13532822 163456.3 127908.9 135932.7 130704.9 118312.7 158690.6
                     DP     Wind mean_aqi     WBGT     Flag
IncNodePurity 125917.9 105497.3 198770.4 120521.3 14779.83
```

are included to account for potential age- and sex-related differences in how weather impacts performance.

Table 5 summarizes the linear regression model. The results show that X.rh (Percent Relative Humidity), SR.W.m2 (Solar Radiation in Watts per square meter), and mean_aqi (Air Quality Index) are statistically significant, and their interactions with age are also significant.

However, we are uncertain whether the relationship is truly linear. Generally, the relationship between weather conditions and marathon performance is likely to be complex and non-linear. To address this, we will use random forest[7], a non-parametric model, to perform variable selection and identify the most important weather parameters affecting marathon performance.

Table 3 confirms that the most important weather parameters are SR.W.m2 (Solar Radiation in Watts per square meter) and mean_aqi (Air Quality Index), which is consistent with the findings from the linear regression model.

Therefore, we will plot the relationship between Mean Adjusted Percentage Current Course Record and both Solar Radiation in Watts per square meter and Air Quality Index to visually illustrate the impact of these two weather parameters on marathon performance.

Figure 7 shows the relationship between Adjusted X.CR (Adjusted Percentage Current Course Record) and solar radiation (measured in Watts per square meter) across different age groups. The plot reveals that higher levels of solar radiation are associated with worse marathon performance, as indicated by higher Adjusted X.CR values. Performance tends to be better when solar radiation is lower, and this trend holds across various age bins.

In Figure 8, the impact of the Air Quality Index (AQI) on Adjusted X.CR is displayed. The plot indicates that as AQI increases, indicating poorer air quality, marathon performance declines. This trend is evident across all age bins, suggesting that poor air quality negatively affects marathon performance regardless of age, though younger athletes seem less affected compared to older runners.

Figure 10 in Appendix shows the relationship between Adjusted X.CR and Wet Bulb Globe Temperature (WBGT) across different age groups. While the trend is similar to previous plots, with performance generally worsening as WBGT increases, the relationship is not as

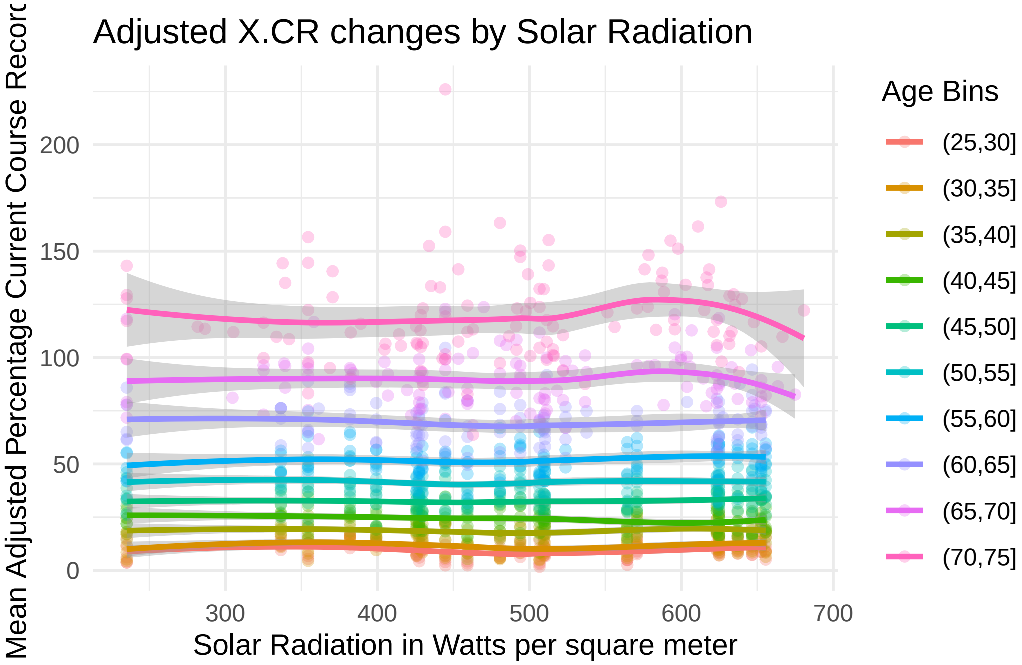


Figure 7

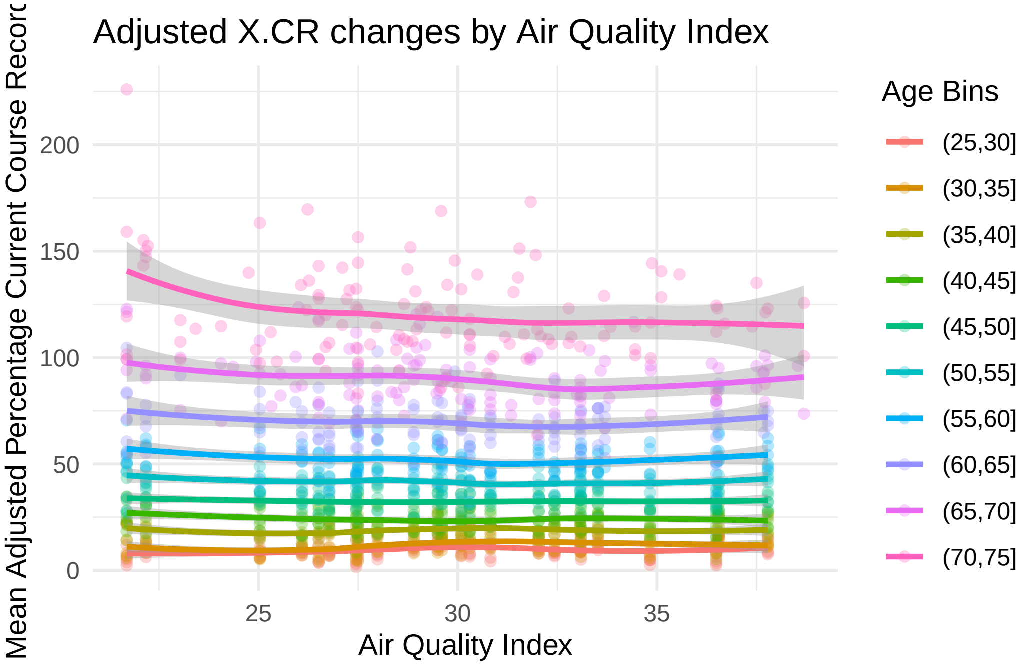


Figure 8

pronounced or significant compared to the impact of other weather variables like solar radiation or air quality.

Conclusion:

Throught the statistical model, we identified that Solar Radiation in Watts per square meter and Air Quality Index are the most important weather parameters affecting marathon performance. The relationship between these weather parameters and marathon performance is complex and non-linear, with performance being better in lower AQI conditions and lower solar radiation. This indicates that air quality and solar radiation have a significant impact on marathon performance.

References

1. Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. Med Sci Sports Exerc, 42(1), 135-41.
2. Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. Medicine and science in sports and exercise, 39(3), 487-493.
3. Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.
4. Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., … & Millet, G. Y. (2022). Sex differences in endurance running. Sports medicine, 52(6), 1235-1257.
5. Yanovich, R., Ketko, I., & Charkoudian, N. (2020). Sex differences in human thermoregulation: relevance for 2020 and beyond. Physiology, 35(3), 177-184.
6. Lloyd S. Least squares quantization in PCM[J]. IEEE transactions on information theory, 1982, 28(2): 129-137.
7. Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.
8. https://aqs.epa.gov/aqsweb/documents/codetables/parameter_classes.html

Appendix

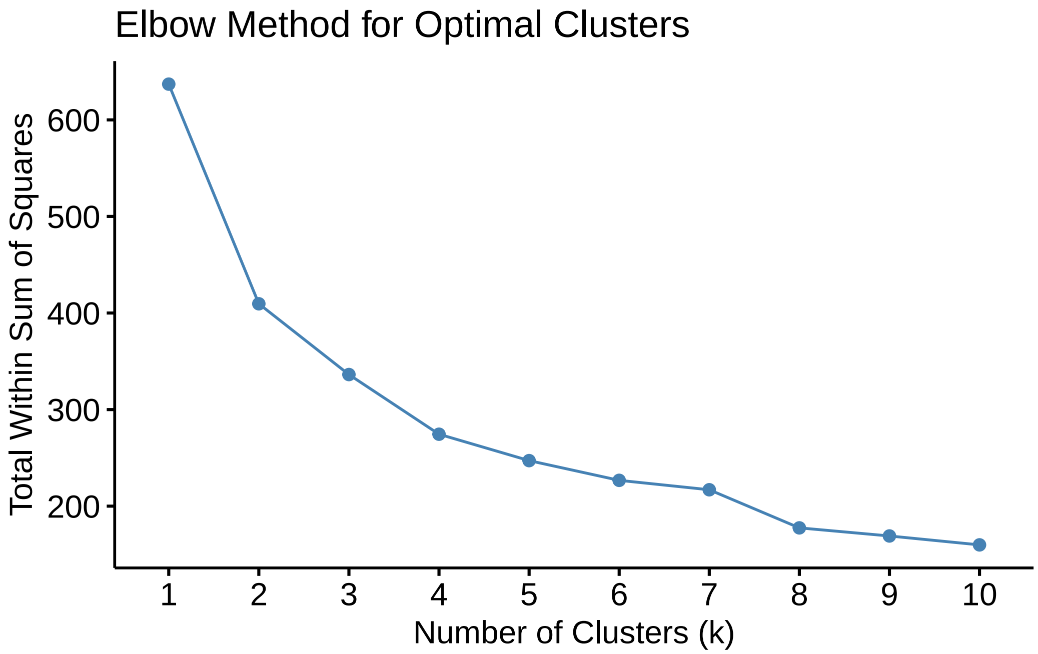


Figure 9

Table 4: Table of Cluster Flag matrix

```
  White Green Yellow Red
1     3    11      0   0
2     1     1      0   0
3     0    10      1   0
4     0     0      1   2
5     7    14      0   0
6    20     0      0   0
7     0     2     10   0
8     0     1      5   3
```

Table 5: Summary table of linear regression model

```
Call:
lm(formula = Adjust_X.CR ~ Age + Gender + Td..C + Tw..C + X.rh +
    Tg..C + SR.W.m2 + DP + Wind + mean_aqi + Age * Gender + Age *
    Td..C + Age * Tw..C + Age * X.rh + Age * Tg..C + Age * SR.W.m2 +
    Age * DP + Age * Wind + Age * mean_aqi + Gender * Td..C +
    Gender * Tw..C + Gender * X.rh + Gender * Tg..C + Gender *
    SR.W.m2 + Gender * DP + Gender * Wind + Gender * mean_aqi,
    data = aim3_filter)

Residuals:
    Min      1Q  Median      3Q     Max
-72.372 -13.719  -3.263   8.999 193.675

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.941e+01  6.701e+00 -13.343  < 2e-16 ***
Age                  2.791e+00  1.255e-01  22.237  < 2e-16 ***
GenderMale          -2.072e+01  4.078e+00  -5.082 3.81e-07 ***
Td..C               -2.635e+00  1.153e+00  -2.285 0.022307 *
Tw..C                3.601e+00  2.420e+00   1.488 0.136738
X.rh                 1.179e-01  3.470e-02   3.397 0.000685 ***
Tg..C               -3.613e-01  2.767e-01  -1.306 0.191747
SR.W.m2              3.561e-02  6.889e-03   5.169 2.40e-07 ***
DP                  -8.568e-01  1.088e+00  -0.787 0.431210
Wind                 3.447e-01  2.465e-01   1.399 0.161919
mean_aqi             3.465e-01  7.689e-02   4.506 6.69e-06 ***
Age:GenderMale       1.372e-01  3.077e-02   4.459 8.34e-06 ***
Age:Td..C            6.820e-02  2.180e-02   3.128 0.001765 **
Age:Tw..C           -7.415e-02  4.580e-02  -1.619 0.105425
Age:X.rh            -2.907e-03  6.501e-04  -4.471 7.87e-06 ***
Age:Tg..C            1.115e-02  5.360e-03   2.080 0.037557 *
Age:SR.W.m2         -1.037e-03  1.311e-04  -7.911 2.84e-15 ***
Age:DP               1.279e-02  2.060e-02   0.621 0.534660
Age:Wind            -7.055e-03  4.607e-03  -1.531 0.125761
Age:mean_aqi        -9.887e-03  1.442e-03  -6.858 7.41e-12 ***
GenderMale:Td..C    -9.211e-01  6.503e-01  -1.416 0.156691
GenderMale:Tw..C     2.468e+00  1.362e+00   1.811 0.070096 .
GenderMale:X.rh      5.809e-02  1.959e-02   2.965 0.003036 **
GenderMale:Tg..C    -1.125e-01  1.556e-01  -0.723 0.469882
GenderMale:SR.W.m2   1.013e-02  3.883e-03   2.608 0.009131 **
GenderMale:DP       -1.154e+00  6.123e-01  -1.885 0.059471 .
GenderMale:Wind      6.005e-02  1.391e-01   0.432 0.666059
GenderMale:mean_aqi  3.856e-02  4.340e-02   0.889 0.374205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.11 on 9326 degrees of freedom
Multiple R-squared:  0.7282,    Adjusted R-squared:  0.7274
F-statistic: 925.4 on 27 and 9326 DF,  p-value: < 2.2e-16
```

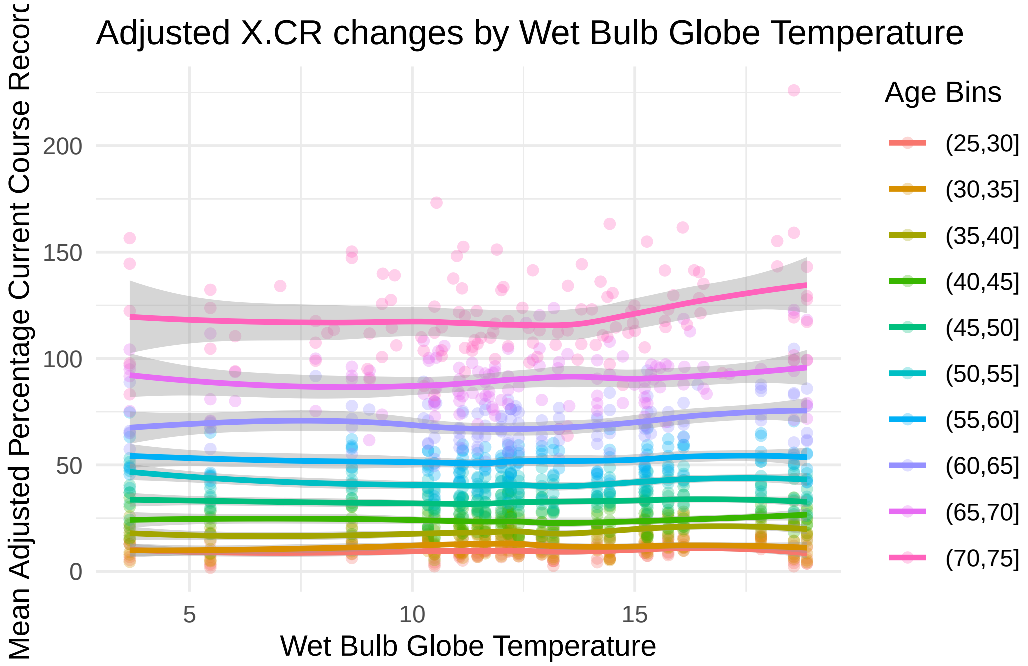


Figure 10

Code Appendix:

```
library(tidyverse)
library(dplyr)
library(lubridate)
library(factoextra)
library(randomForest)
library(kableExtra)
library(gtsummary)
library(ggplot2)
library(reshape2)

project1 <- read.csv("project1.csv")
course_record <- read.csv("course_record.csv")
marathon_dates <- read.csv("marathon_dates.csv")
aqi_values <- read.csv("aqi_values.csv")
# Load the dataset
colnames(project1)[c(1,3,5)] <- c("Race","Gender","Age")
```

```
# Create a summary table
summary_table <- project1 %>%
  group_by(Race) %>%
  summarize(
    Avg_Age = mean(Age, na.rm = TRUE),
    Avg_CR = mean(X.CR, na.rm = TRUE),
    Avg_Td = mean(Td..C, na.rm = TRUE),
    Avg_Tw = mean(Tw..C, na.rm = TRUE),
    Avg_rh = mean(X.rh, na.rm = TRUE),
    Avg_Tg = mean(Tg..C, na.rm = TRUE),
    Avg_SR = mean(SR.W.m2, na.rm = TRUE),
    Avg_DP = mean(DP, na.rm = TRUE),
    Avg_Wind = mean(Wind, na.rm = TRUE),
    Avg_WBGT = mean(WBGT, na.rm = TRUE)
  )

# View the summary table
print(summary_table)
project1 <- na.omit(project1)

# rename the column names and variables
course_record$Race <- recode(course_record$Race,
                             "B" = "BOS", "C" = "CHI", "NY" = "NYC", "TC" = "TCM", "D" = "DL
course_record$Gender <- recode(course_record$Gender, "F" = "Female", "M" = "Male")
course_record$CR <- as.numeric(hms(course_record$CR))

colnames(project1)[c(1,3,5)] <- c("Race","Gender","Age")
project1$Race <- recode(project1$Race,"0" = "BOS", "1" = "CHI", "2" = "NYC", "3" = "TCM", "4
project1$Gender <- recode(project1$Gender, "0" = "Female", "1" = "Male")

# calculate the mean record
course_record_mean <- course_record %>%
  group_by(Race, Year, Gender) %>%
  summarise(Mean_Record = mean(CR),.groups = 'drop') %>%
  ungroup()
course_record_mean$CR <- course_record$CR

# Air Quality Data Processing
aqi_values_mean <-
  aqi_values[!is.na(aqi_values$aqi),] %>%
  group_by(date_local, marathon) %>%
  filter(sample_duration == "8-HR RUN AVG BEGIN HOUR") %>%
  summarise(mean_aqi = mean(aqi), .groups = 'drop')
aqi_values_mean$date_local <- year(ymd(aqi_values_mean$date_local))
aqi_values_mean$marathon <- recode(aqi_values_mean$marathon,"Boston" = "BOS", "Chicago" = "C
colnames(aqi_values_mean)[c(1,2)] <- c("Year","Race")

# Merge the datasets on 'Race' and 'Year'
project_merged <- merge(project1, course_record_mean, by = c("Race","Year","Gender"))

# merge air data on Aqi
project_merged <- left_join(project_merged, aqi_values_mean,by = c('Year', 'Race'))

# Calculate the adjusted CR
project1_adjust <- project_merged %>%
  mutate(Adjust_X.CR = X.CR * (CR + 1) / Mean_Record)

# project1_adjust %>%
#   select(X.CR, Adjust_X.CR, mean_aqi) %>%
#   summary()
# calculate the mean,max,min of Adjust_X.CR
project_1_aim1 <- project1_adjust %>%
  group_by(Age,Gender) %>%
  summarise(mean_Adjust_X.CR = mean(Adjust_X.CR),min_Adjust_X.CR = min(Adjust_X.CR),max_Adju
            .groups = "drop")
ggplot(project_1_aim1, aes(x = as.factor(Age), y = mean_Adjust_X.CR,
                           color = as.factor(Gender),
                           linetype = as.factor(Gender),
                           group = Gender)) +
  geom_line(linewidth = 0.5) +  # Line for the mean
  geom_point(size = 0.5) +  # Points for the mean values
  geom_errorbar(aes(ymin = min_Adjust_X.CR, ymax = max_Adjust_X.CR), width = 1) +
  labs(x = "Age(Years)", y = "mean adjust CR", fill = "Sex(0=F, 1=M)",
       title = "Mean Adjusted Percentage Course Record with error bar by Age and Sex") +
  theme_minimal() +
  # Customizing the legend labels for gender
  scale_color_discrete(labels = c("Female", "Male"), name = 'Sex') +
  scale_linetype_discrete(labels = c("Female", "Male"), name = 'Sex') +
  # Set x-axis breaks for better readability
  scale_x_discrete(breaks = seq(0, 100, by = 10))
ggplot(filter(project_1_aim1,20<=Age & Age<=70), aes(x = as.factor(Age), y = mean_Adjust_X.C
                           color = as.factor(Gender),
                           linetype = as.factor(Gender),
                           group = Gender)) +
  geom_smooth(linewidth = 1,se = F) +
  geom_point(size = 0.5) +  # Points for the mean values
  # geom_errorbar(aes(ymin = min_Adjust_X.CR, ymax = max_Adjust_X.CR), width = 1) +  # Error
  labs(x = "Age(Years)", y = "mean adjust CR", fill = "Sex(0=F, 1=M)",
       title = "Fitted curve of Mean Adjusted Percentage Current Course Record by Age and Se
  theme_minimal() +
  # Customizing the legend labels for gender
  scale_color_discrete(labels = c("Female", "Male"), name = 'Sex') +
  scale_linetype_discrete(labels = c("Female", "Male"), name = 'Sex') +
  scale_x_discrete(breaks = seq(0, 100, by = 10))
  # Set x-axis breaks for better readability
# process the data to get environmental conditions
aim2_grouped <-
  project1_adjust %>%
  group_by(Race, Year) %>%
  summarise(Flag = ifelse(length(unique(Flag)) == 1, unique(Flag), NA),
            Td..C = ifelse(length(unique(Td..C)) == 1, unique(Td..C), NA),
            Tw..C = ifelse(length(unique(Tw..C)) == 1, unique(Tw..C), NA),
            X.rh = ifelse(length(unique(X.rh)) == 1, unique(X.rh), NA),
            Tg..C = ifelse(length(unique(Tg..C)) == 1, unique(Tg..C), NA),
            SR.W.m2 = ifelse(length(unique(SR.W.m2)) == 1, unique(SR.W.m2), NA),
            DP = ifelse(length(unique(DP)) == 1, unique(DP), NA),
            Wind = ifelse(length(unique(Wind)) == 1, unique(Wind), NA),
            WBGT = ifelse(length(unique(WBGT)) == 1, unique(WBGT), NA),
            .groups = 'drop')
aim2_cluster <- aim2_grouped %>% select(-c(Race:Flag,WBGT))

# Standardize the data
aim2_cluster_scaled <- scale(aim2_cluster)

# Apply KMeans clustering with 4 clusters
set.seed(2550)  # For reproducibility
kmeans_result <- kmeans(aim2_cluster_scaled, centers = 8)

# Add the cluster assignments to the original data
aim2_grouped$Cluster <- kmeans_result$cluster

# Reorder the columns: White, Green, Yellow, Red
cluster_flag_matrix <- table(aim2_grouped$Cluster, aim2_grouped$Flag)[, c("White", "Green", "
aim2_select <- aim2_grouped %>%
  filter(Cluster %in% c(3,4,6,7))
```

```r
# Define custom colors for the Flag
custom_colors <- c("Green" = "green", "White" = "blue", "Yellow" = "yellow", "Red" = "red")
aim2_select$Cluster <- as.factor(aim2_select$Cluster)

# Create the scatter plot with custom colors for Flag
ggplot(aim2_select, aes(x = Td..C, y = Tw..C, color = Flag, shape = Cluster)) +
  geom_point(size = 3) +
  scale_color_manual(values = custom_colors) +   # Apply custom colors
  labs(x = "Td (°C) - Dry Bulb Temperature",
       y = "Tw (°C) - Wet Bulb Temperature",
       title = "KMeans Clustering of environmental conditions with Flag Colors") +
  theme_minimal()
aim2_merge <-
  merge(aim2_grouped, project1_adjust, by = c("Race","Year")) %>%
  group_by(Cluster, Gender, Age) %>%
  summarise(Mean_Adjust_X.CR = mean(Adjust_X.CR), .groups = 'drop')

aim2_merge_filtered <- aim2_merge %>%
  filter(Age >= 20 & Age <= 80) %>%
  filter(Cluster %in% c(3,4,6,7))

aim2_merge_filtered$Cluster <- recode(aim2_merge_filtered$Cluster,
                             "3" = "Normal_1",
                             "4" = "Bad",
                             "6" = "Good",
                             "7" = "Normal_2")

# Plot using ggplot with facet_wrap and restricted age range
ggplot(aim2_merge_filtered,
       aes(x = Age, y = Mean_Adjust_X.CR, color = factor(Cluster))) +
  geom_line() +
  facet_wrap(~ Gender) +
  labs(title = "Mean Adjusted Percentage Current Course Record by Gender, Weather (Age 20-80)
       x = "Age",
       y = "Mean_Adjust_X.CR",
       color = "Cluster") +
  theme_minimal() +
  scale_color_discrete(name = "Weather Condition")
# Calculate the derivative
aim2_merge$derivative <- c(NA,abs(diff(aim2_merge$Mean_Adjust_X.CR)))

aim2_merge_diff <- aim2_merge %>%
  filter(Age >= 25 & Age <= 70) %>%
  filter(Cluster %in% c(3,4,6,7))

aim2_merge_diff$Cluster <- recode(aim2_merge_diff$Cluster,
                             "3" = "Normal_1",
                             "4" = "Bad",
                             "6" = "Good",
                             "7" = "Normal_2")

# Create a single plot using facet_wrap to separate Gender
ggplot(aim2_merge_diff, aes(x = Age, y = derivative, color = factor(Cluster))) +
  geom_smooth(linewidth = 0.5,se = F) +
  facet_wrap(~ Gender) +
    labs(title = "Difference of Adjusted CR by Gender, Weather (Age 25-70)", x = "Age", y = "
  theme_minimal() +
  scale_color_discrete(name = "Weather Condition")
# Calculate correlation matrix
cor_matrix <- cor(project1_adjust[, c('Age', 'WBGT', 'Td..C', 'Tw..C', "X.rh", "Tg..C", 'SR.

# Reshape the data for ggplot
cor_melt <- melt(as.matrix(cor_matrix))

# Plot the heatmap
ggplot(data = cor_melt, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0,
                      limit = c(-1, 1), space = "Lab",
                      name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  labs(x = "Variables", y = "Variables", title = "Heatmap of Correlation Matrix")
# Calculate correlations by age
aim3_corr <-
  project1_adjust %>%
  filter(Age >= 20 & Age <= 80) %>%
  select(-Race, -Year, -X.CR, -Flag, -Gender, -CR, -Mean_Record) %>%
  group_by(Age) %>%
  summarise(across(everything(), ~ cor(.x, Adjust_X.CR, use = "complete.obs"), .names = "{.c
  select(-Adjust_X.CR)

# Reshape the data for plotting
correlation_long <- pivot_longer(aim3_corr, -Age, names_to = "Variable", values_to = "Correla
correlation_long$Variable <- recode(correlation_long$Variable,
                             "DP" = "Dew Point in Celsius)",
                             "Td..C" " = "Bulb Temperaturein Celsius",
                             "Tw..C" = "Wet Bulb Temperature in Celsius",
                             "X.rh" = "Percent Relative Humidity",
                             "Tg..C" = "Black Globe Temperature in Celsius",
                             "Td..C" = "Dry Bulb Temperature in Celsius",
                             "SR.W.m2" = "Solar Radiation in Watts per square meter",
                             "Wind" = "Wind Speed in km/h",
                             "mean_aqi" = "Air Quality Index",
                             "WBGT" = "Wet Bulb Globe Temperature")

# Plot the data
ggplot(correlation_long, aes(x = Age, y = Correlation, color = Variable, group = Variable)) +
  #geom_line() +
  geom_point(size = 0.5) +
  geom_smooth(se = F) +
  labs(title = "Correlation between Adjust_X.CR and Other Weather Variables by Age", x = "Ag
  theme_minimal() +
  theme(legend.position = "right") +
  scale_color_discrete(name = "Weather Variables")
aim3_filter <- project1_adjust %>%
  filter(Age >= 25 & Age <= 80)

# lm_aim2 <- lm(Adjust_X.CR ~ Age + Gender + Td..C + Tw..C + X.rh + Tg..C + SR.W.m2 + DP + Wi

lm_aim3 <- lm(Adjust_X.CR ~ Age + Gender + Td..C + Tw..C + X.rh + Tg..C + SR.W.m2 + DP + Win
rf_model <- randomForest(Adjust_X.CR ~ Age + Gender + Td..C + Tw..C + X.rh + Tg..C + SR.W.m2

importance(rf_model) %>%t()
aim3_SR.W.m2 <- project1_adjust %>%
  filter(Age > 25 & Age <= 75) %>%
  group_by(Age, Year) %>%
  summarise(Adjust_X.CR = mean(Adjust_X.CR), SR.W.m2 = mean(SR.W.m2),
            .groups = 'drop') %>%
  mutate(Age_Binned = cut(Age, breaks = seq(25, 75, by = 5)))

# Set up the plot
ggplot(aim3_SR.W.m2, aes(x = SR.W.m2, y = Adjust_X.CR, color = Age_Binned)) +
  geom_point(alpha = 0.3) +  # Scatter points
  geom_smooth(se = 0.2) +  # Polynomial fit (degree 2)
  labs(x = "Solar Radiation in Watts per square meter", y = "Mean Adjusted Percentage Curren
       title = "Adjusted X.CR changes by Solar Radiation",
       color = "Age Bins") +
  theme_minimal() +
  theme(legend.position = "right")
aim3_aqi <- project1_adjust %>%
  filter(Age > 25 & Age <= 75) %>%
  filter(mean_aqi <= 50) %>%
  group_by(Age, Year) %>%
  summarise(Adjust_X.CR = mean(Adjust_X.CR), mean_aqi = mean(mean_aqi),
            .groups = 'drop') %>%
  mutate(Age_Binned = cut(Age, breaks = seq(25, 75, by = 5)))

# Set up the plot
ggplot(aim3_aqi, aes(x = mean_aqi, y = Adjust_X.CR, color = Age_Binned)) +
  geom_point(alpha = 0.3) +  # Scatter points
  geom_smooth(se = 0.2) +  # Polynomial fit (degree 2)
  labs(x = "Air Quality Index", y = "Mean Adjusted Percentage Current Course Record",
       title = "Adjusted X.CR changes by Air Quality Index",
       color = "Age Bins") +
  theme_minimal() +
  theme(legend.position = "right")
# Set seed for reproducibility
set.seed(2550)

# Perform K-means clustering with different numbers of clusters (k) and plot the elbow curve
fviz_nbclust(aim2_cluster_scaled, kmeans, method = "wss") +
  labs(title = "Elbow Method for Optimal Clusters",
       x = "Number of Clusters (k)",
       y = "Total Within Sum of Squares")
print(cluster_flag_matrix)
summary(lm_aim3)
aim3_WBGT <- project1_adjust %>%
  filter(Age > 25 & Age <= 75) %>%
  group_by(Age, Year) %>%
  summarise(Adjust_X.CR = mean(Adjust_X.CR), WBGT = mean(WBGT),
            .groups = 'drop') %>%
  mutate(Age_Binned = cut(Age, breaks = seq(25, 75, by = 5)))

# Set up the plot
ggplot(aim3_WBGT, aes(x = WBGT, y = Adjust_X.CR, color = Age_Binned)) +
  geom_point(alpha = 0.3) +  # Scatter points
  geom_smooth(se = 0.2) +  # Polynomial fit (degree 2)
  labs(x = "Wet Bulb Globe Temperature", y = "Mean Adjusted Percentage Current Course Record"
       title = "Adjusted X.CR changes by Wet Bulb Globe Temperature",
       color = "Age Bins") +
  theme_minimal() +
  theme(legend.position = "right")
```