

Project 1 EDA

Diahmin Hawkins dlh2166@columbia.edu

9/19/2024

#Data

```
library(readr)
marathon_dates <- read_csv("marathon_dates.csv")

## Rows: 98 Columns: 3
## -- Column specification -----
## Delimiter: ","
## chr  (1): marathon
## dbl  (1): year
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
course_record <- read_csv("course_record.csv")

## Rows: 194 Columns: 4
## -- Column specification -----
## Delimiter: ","
## chr  (2): Race, Gender
## dbl  (1): Year
## time (1): CR
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
aqi_values <- read_csv("aqi_values.csv")

## Rows: 10451 Columns: 11
## -- Column specification -----
## Delimiter: ","
## chr  (5): county_code, site_number, units_of_measure, sample_duration, marathon
## dbl  (5): cbsa_code, state_code, parameter_code, aqi, arithmetic_mean
## date (1): date_local
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
project1 <- read_csv("project1.csv")

## Rows: 11564 Columns: 14
## -- Column specification -----
```

```

## Delimiter: ","
## chr (1): Flag
## dbl (13): Race (0=Boston, 1=Chicago, 2=NYC, 3=TC, 4=D), Year, Sex (0=F, 1=M)...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

#Packages
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr 1.1.4 v stringr 1.5.1
## v forcats 1.0.0 v tibble 3.2.1
## v ggplot2 3.5.1 v tidyr 1.3.1
## v purrr 1.0.2

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidyr)
library(dplyr)
library(naniar)
library(visdat)
library(kableExtra)

##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows

library(knitr)
library(gridExtra)

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine

library(ggribbles)

# head(course_record)
# head(marathon_dates)
# head(project1) # Change project 1 column name to gender
# marathon_dates

```

#Course Record Data Management

```
course_record<- course_record%>%
  mutate(Race_Seconds= as.numeric(hms(course_record$CR)))

#Change column names from Sex... to Gender to match project1 dataset
colnames(course_record)[colnames(course_record) == "Sex (0=F, 1=M)"] ="Gender"

# Mutate the Race names from numbers to the marathon cities for better understanding

# Change the Race variable to a character variable
course_record <- course_record %>%
  mutate(Race = as.character(Race))

head(course_record)
```

```
## # A tibble: 6 x 5
##   Race   Year CR      Gender Race_Seconds
##   <chr> <dbl> <time>   <chr>      <dbl>
## 1 B     2016 02:03:02 M          7382
## 2 B     2015 02:03:02 M          7382
## 3 B     2014 02:03:02 M          7382
## 4 B     2013 02:03:02 M          7382
## 5 B     2012 02:03:02 M          7382
## 6 B     2011 02:05:52 M          7552
```

Project 1 Data Management

#Change colnames for a more readable and understanding approach

```
colnames(project1)[colnames(project1) == "Sex (0=F, 1=M)"] ="Gender"
colnames(project1)[colnames(project1) == "Race (0=Boston, 1=Chicago, 2=NYC, 3=TC, 4=D)"] ="Race"
colnames(project1)[colnames(project1) == "Td, C"] ="Dry bulb Temp C"
colnames(project1)[colnames(project1) == "Tw, C"] ="Wet bulb Temp C"
colnames(project1)[colnames(project1) == "%rh"] ="Percent Relative Humidity"
colnames(project1)[colnames(project1) == "Tg, C"] ="Black Globe Temp C"
colnames(project1)[colnames(project1) == "SR W/m2"] ="Solar Radiation"
colnames(project1)[colnames(project1) == "DP"] ="Dew Point in C"
colnames(project1)[colnames(project1) == "Wind"] ="Wind Speed"
colnames(project1)[colnames(project1) == "WBGT"] ="Wet Bulb Globe Temp"
colnames(project1)[colnames(project1) == "%CR"] ="Percent CR"
```

#change gender if 0 to F to represent female

```
project1$Gender<-ifelse(project1$Gender== "0","F","M") #change gender if 0 to F to represent female el.
```

Mutate the Race names from numbers to the marathon cities for better understanding

```
project1 <- project1 %>%
  mutate(Race = case_when(
    Race == "0" ~ "B",
    Race == "1" ~ "C",
    Race == "2" ~ "NY",
    Race == "3" ~ "TC",
```

```

Race == "4" ~ "D"))

#Change the Race variable to a character variable
project1 <- project1 %>%
  mutate(Race = as.character(Race))

# Merge the two dataframes
course_record_project1<-left_join(project1,course_record,
                                   by= c("Race","Gender", "Year"))

# Change the Course Percentage %CR into course seconds
course_record_project1 <- course_record_project1 %>%
  mutate(Runtimes = Race_Seconds * (1 + (`Percent CR` / 100)))

head(course_record_project1)

## # A tibble: 6 x 17
##   Race   Year Gender Flag   `Age (yr)` `Percent CR` `Dry bulb Temp C`
##   <chr> <dbl> <chr> <chr>     <dbl>         <dbl>         <dbl>
## 1 B     2016 M     Green     18          35.7          13.8
## 2 B     2016 M     Green     19          39.3          13.8
## 3 B     2016 M     Green     20          15.7          13.8
## 4 B     2016 M     Green     21           7.90         13.8
## 5 B     2016 M     Green     22          24.7          13.8
## 6 B     2016 M     Green     23          10.3          13.8
## # i 10 more variables: `Wet bulb Temp C` <dbl>,
## #   `Percent Relative Humidity` <dbl>, `Black Globe Temp C` <dbl>,
## #   `Solar Radiation` <dbl>, `Dew Point in C` <dbl>, `Wind Speed` <dbl>,
## #   `Wet Bulb Globe Temp` <dbl>, CR <time>, Race_Seconds <dbl>, Runtimes <dbl>

head(project1)

## # A tibble: 6 x 14
##   Race   Year Gender Flag   `Age (yr)` `Percent CR` `Dry bulb Temp C`
##   <chr> <dbl> <chr> <chr>     <dbl>         <dbl>         <dbl>
## 1 B     2016 M     Green     18          35.7          13.8
## 2 B     2016 M     Green     19          39.3          13.8
## 3 B     2016 M     Green     20          15.7          13.8
## 4 B     2016 M     Green     21           7.90         13.8
## 5 B     2016 M     Green     22          24.7          13.8
## 6 B     2016 M     Green     23          10.3          13.8
## # i 7 more variables: `Wet bulb Temp C` <dbl>,
## #   `Percent Relative Humidity` <dbl>, `Black Globe Temp C` <dbl>,
## #   `Solar Radiation` <dbl>, `Dew Point in C` <dbl>, `Wind Speed` <dbl>,
## #   `Wet Bulb Globe Temp` <dbl>

head(course_record)

## # A tibble: 6 x 5
##   Race   Year CR      Gender Race_Seconds
##   <chr> <dbl> <time>  <chr>         <dbl>
## 1 B     2016 02:03:02 M           7382
## 2 B     2015 02:03:02 M           7382

```

```
## 3 B      2014 02:03:02 M      7382
## 4 B      2013 02:03:02 M      7382
## 5 B      2012 02:03:02 M      7382
## 6 B      2011 02:05:52 M      7552
```

Missing Data Attributes

Possibly Meet Completely at random because the the course_record wasn't dependent on the project 1 data. They carried some of the same attributes and shared covariates the we can't say one is dependent of the other.

```
#Get the sum on NA's to get assumptions (MCAR)
```

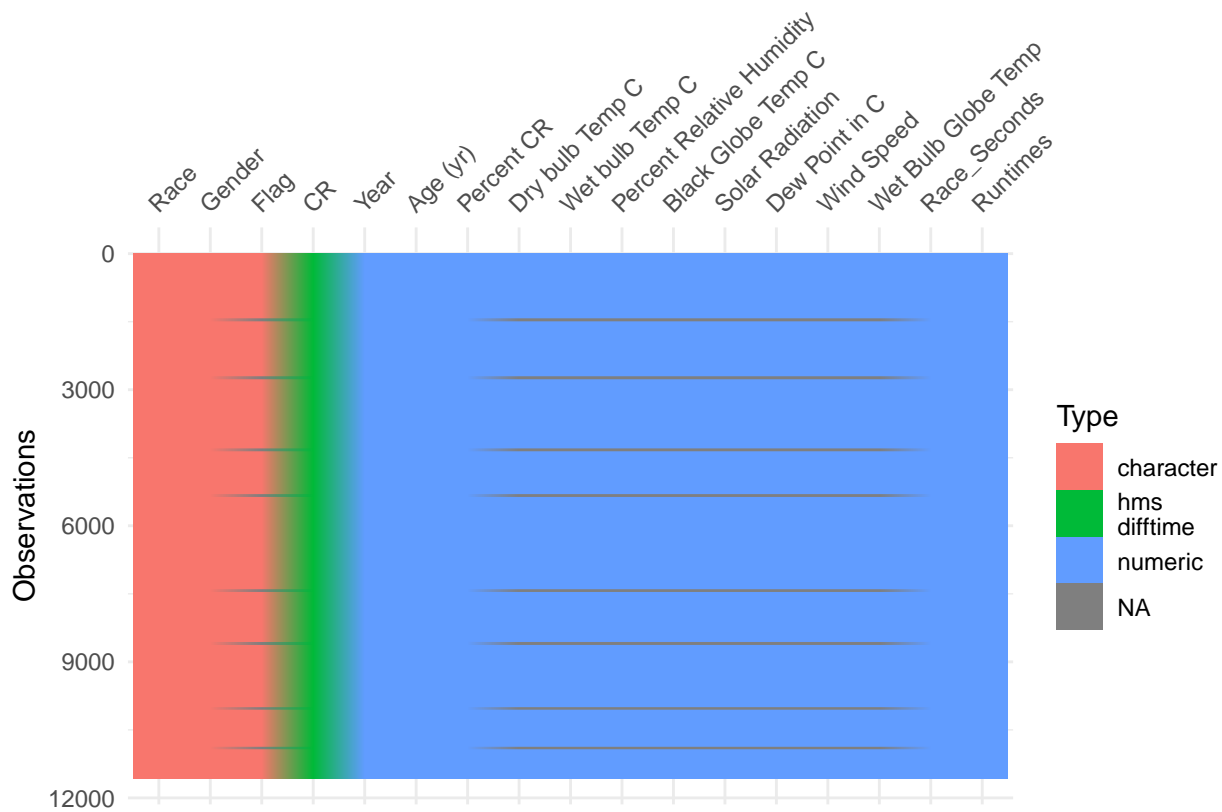
```
sum(is.na(course_record_project1))
```

```
## [1] 4419
```

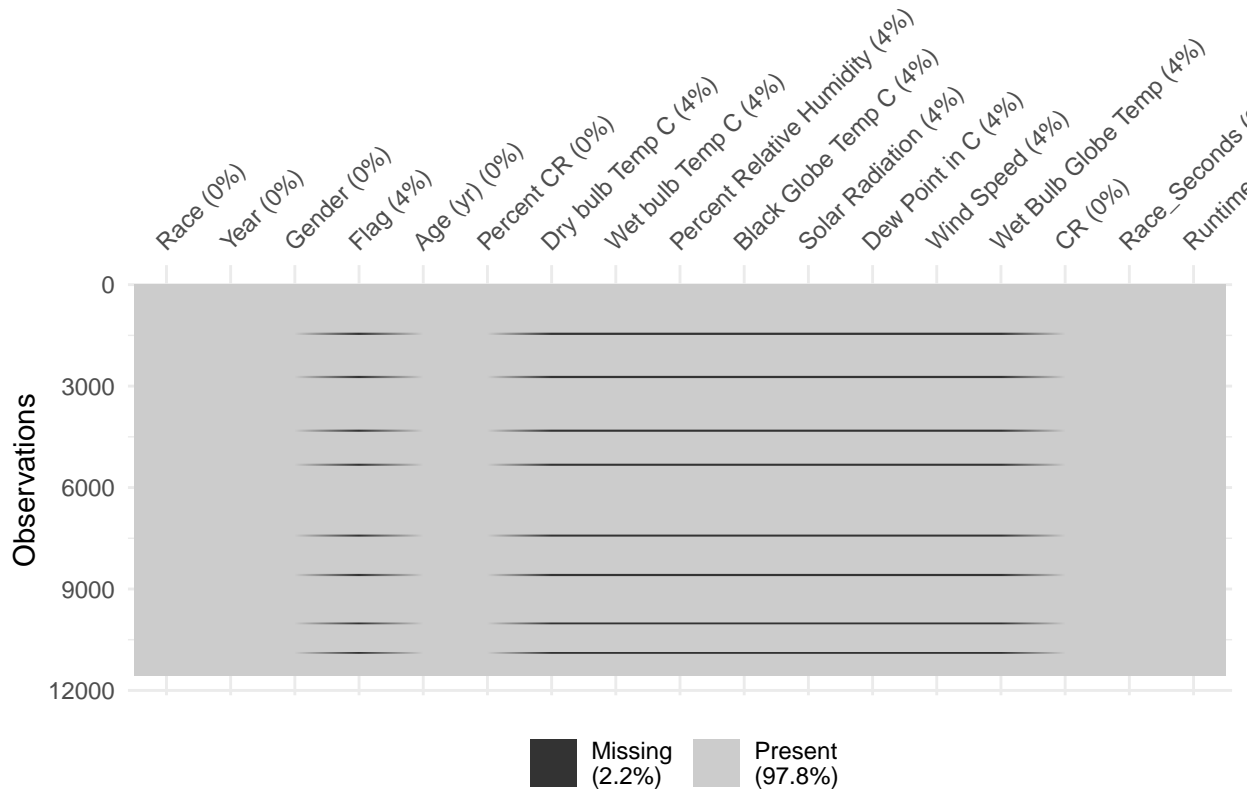
```
str(course_record_project1)
```

```
## tibble [11,564 x 17] (S3: tbl_df/tbl/data.frame)
## $ Race      : chr [1:11564] "B" "B" "B" "B" ...
## $ Year      : num [1:11564] 2016 2016 2016 2016 2016 ...
## $ Gender    : chr [1:11564] "M" "M" "M" "M" ...
## $ Flag      : chr [1:11564] "Green" "Green" "Green" "Green" ...
## $ Age (yr)   : num [1:11564] 18 19 20 21 22 23 24 25 26 27 ...
## $ Percent CR : num [1:11564] 35.7 39.3 15.7 7.9 24.7 ...
## $ Dry bulb Temp C : num [1:11564] 13.8 13.8 13.8 13.8 13.8 ...
## $ Wet bulb Temp C : num [1:11564] 8.23 8.23 8.23 8.23 8.23 ...
## $ Percent Relative Humidity: num [1:11564] 45.6 45.6 45.6 45.6 45.6 ...
## $ Black Globe Temp C : num [1:11564] 28 28 28 28 28 ...
## $ Solar Radiation : num [1:11564] 766 766 766 766 766 ...
## $ Dew Point in C : num [1:11564] 2.23 2.23 2.23 2.23 2.23 ...
## $ Wind Speed  : num [1:11564] 12.7 12.7 12.7 12.7 12.7 ...
## $ Wet Bulb Globe Temp : num [1:11564] 12.7 12.7 12.7 12.7 12.7 ...
## $ CR         : 'hms' num [1:11564] 02:03:02 02:03:02 02:03:02 02:03:02 ...
## ..- attr(*, "units")= chr "secs"
## $ Race_Seconds : num [1:11564] 7382 7382 7382 7382 7382 ...
## $ Runtimes     : num [1:11564] 10019 10280 8541 7965 9203 ...
```

```
course_record_project1%>% vis_dat()
```



```
vis_miss(course_record_project1)
```



```
course_record_project1%>% glimpse()
```

```
## Rows: 11,564
## Columns: 17
## $ Race          <chr> "B", "B", "B", "B", "B", "B", "B", "B", "B~
## $ Year          <dbl> 2016, 2016, 2016, 2016, 2016, 2016, 2016, ~
## $ Gender        <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M~
## $ Flag          <chr> "Green", "Green", "Green", "Green", "Green~
## $ `Age (yr)`    <dbl> 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28~
## $ `Percent CR`  <dbl> 35.722027, 39.257654, 15.700352, 7.897589, ~
## $ `Dry bulb Temp C` <dbl> 13.82857, 13.82857, 13.82857, 13.82857, 13~
## $ `Wet bulb Temp C` <dbl> 8.228571, 8.228571, 8.228571, 8.228571, 8.~
## $ `Percent Relative Humidity` <dbl> 45.57143, 45.57143, 45.57143, 45.57143, 45~
## $ `Black Globe Temp C` <dbl> 27.97143, 27.97143, 27.97143, 27.97143, 27~
## $ `Solar Radiation` <dbl> 765.5857, 765.5857, 765.5857, 765.5857, 76~
## $ `Dew Point in C` <dbl> 2.228571, 2.228571, 2.228571, 2.228571, 2.~
## $ `Wind Speed`  <dbl> 12.71429, 12.71429, 12.71429, 12.71429, 12~
## $ `Wet Bulb Globe Temp` <dbl> 12.73714, 12.73714, 12.73714, 12.73714, 12~
## $ CR            <time> 02:03:02, 02:03:02, 02:03:02, 02:03:02, 0~
## $ Race_Seconds  <dbl> 7382, 7382, 7382, 7382, 7382, 7382, 7382, ~
## $ Runtimes      <dbl> 10019, 10280, 8541, 7965, 9203, 8145, 8949~
```

```
#Get the sum of Missing Data for each column
```

```
Missing_Data<- sapply(course_record_project1, function(x) sum(is.na(x)))
```

```
# Convert to dataframe
```

```
Missing_Data_df <- data.frame(ColumnNames = names(Missing_Data), `Missing Data` = Missing_Data)
```

```
# Set names for the dataframe columns i
```

```
names(Missing_Data_df) <- c("Variables", "Missing Data")
```

```
#Make the Missing Data to a dataframe
```

```
as.data.frame(Missing_Data)
```

```
##              Missing_Data
## Race              0
## Year              0
## Gender            0
## Flag             491
## Age (yr)          0
## Percent CR        0
## Dry bulb Temp C   491
## Wet bulb Temp C   491
## Percent Relative Humidity 491
## Black Globe Temp C 491
## Solar Radiation   491
## Dew Point in C    491
## Wind Speed        491
## Wet Bulb Globe Temp 491
## CR                0
## Race_Seconds      0
## Runtimes          0
```

```
#Create Variable Description Dataframe
```

```
Variables_dataframe<- data_frame(
```

```

Variables= c("Race", "Year", "Gender", "Flag", "Age (yr)",
             "Percent CR", "Dry bulb Temp C", "Wet bulb Temp C",
             "Percent Relative Humidity", "Black Globe Temp C", "Solar Radiation",
             "Dew Point in C", "Wind Speed", "Wet Bulb Globe Temp", "CR",
             "Race_Seconds", "Runtimes", "age_ranges"),
Type= c("Character", "Numeric", "Character", "Character", "Numeric",
        "Numeric", "Numeric", "Numeric", "Numeric", "Numeric", "Numeric", "Numeric",
        "Numeric", "Numeric", "HMS/Numeric", "Numeric", "Numeric", "Categorical"),
Description= c("Race represents the marathons the participants competed, including the B=Boston Marathon,
               C= Chicago Marathon, NY= New York City Marathon, T= Twin Cities Marathon (Minneapolis, MN),
               D= Grandma's Marathon (Duluth, MN).",
               "Years represented in the dataset ranging from 1993-2016.",
               "Gender is represented by F= Female and M= Male.",
               "Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C,
               Red= WBGT >23-28C, and Black= WBGT > 28C",
               "Age (yr) represents the ages of the participants.",
               "Percent CR is the percent off current course record by gender.",
               "Dry bulb Temp Celcius is the air temperature without taking into account of the humidity
               moisture.",
               "Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity.",
               "Percent Relative Humidity how much moisture is in the air compared to the maximum amount.",
               "Black Globe Temp Celcius indicates how hot it feels in direct sunlight. It considers temperature,
               humidity, and solar radiation.",
               "Solar Radiation in Watts per meter squared is the energy emitted by the sun, which translates to
               temperature.",
               "Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure and
               humidity).",
               "Wind Speed in Km/hr.",
               "Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar
               radiation.",
               "CR is the course record for each marathon.",
               "Race_Seconds is the course record measured in seconds.",
               "Runtimes is the converted gender percentage into seconds. This represents the marathon time in
               seconds.",
               "Age_ranges are the age groups.")
)

```

```

## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```

```

# Make Missing Data df Variables a character
Missing_Data_df$Variables <- as.character(Missing_Data_df$Variables)

# Make Variables dataframe Variables a character
Variables_dataframe$Variables <- as.character(Variables_dataframe$Variables)

# Merge the two dataframes
merged_df <- merge(Missing_Data_df, Variables_dataframe, by = "Variables", all = TRUE)

# Create kable_Extra table for the merged_df
merged_df %>%
  kbl(caption = "Marathon Runners' Data Description") %>%
  kable_classic(full_width = F, html_font = "Cambria", font_size = 12)

```

```

# Create the table with kable and customize with kableExtra
kable(merged_df, "latex", booktabs = TRUE, caption = "Marathon Runners' Data Description") %>%
  kable_styling(latex_options = c("striped", "scale_down")) %>%

```


Table 1: Marathon Runners' Data Description

Variables	Missing Data	Type	Description
Age (yr)	0	Numeric	Age (yr) represents the ages of the participants.
age_ranges	NA	Categorical	Age_ranges are the age groups.
Black Globe Temp C	491	Numeric	Black Globe Temp Celcius indicates how hot it feels in direct sunlight. It considers temperature, humidity, wind speed, sun angle, and cloud cover to provide a holistic view of the stress placed on the body in hot environments.
CR	0	HMS/Numeric	CR is the course record for each marathon.
Dew Point in C	491	Numeric	Dew Point in Celcius is the temperature threshold for condensation.
Dry bulb Temp C	491	Numeric	Dry bulb Temp Celcius is the air temperature.
Flag	491	Character	Flag WBGT Thresholds. White= WBGT Thresholds.
Gender	0	Character	Gender is represented by F= Female and M= Male.
Percent CR	0	Numeric	Percent CR is the percent off current course record.
Percent Relative Humidity	491	Numeric	Percent Relative Humidity how much moisture is in the air.
Race	0	Character	Race represents the marathons the participants ran.
Race_Seconds	0	Numeric	Race_Seconds is the course record measured in seconds.
Runtimes	0	Numeric	Runtimes is the converted gender percentage.
Solar Radiation	491	Numeric	Solar Radiation in Watts per meter square.
Wet Bulb Globe Temp	491	Numeric	Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar radiation.
Wet bulb Temp C	491	Numeric	Wet bulb Temp Celcius is a measure of temperature and humidity.
Wind Speed	491	Numeric	Wind Speed in Km/hr.
Year	0	Numeric	Years represented in the dataset ranging from 2010 to 2019.

```

column_spec(1, width = "3cm") %>%
column_spec(2, width = "2cm") %>%
column_spec(3, width = "2cm") %>%
column_spec(4, width = "8cm")

# Assuming 'merged_df' is already loaded and contains your data
# Create the table using kable and customize with kableExtra

kable(merged_df, "latex", booktabs = TRUE, longtable = TRUE, caption = "Marathon Runners' Data Description",
      kable_styling(latex_options = c("striped", "scale_down", "hold_position"), full_width = FALSE) %>%
  column_spec(1, width = "3cm") %>% # Adjust width based on content
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "3cm") %>%
  column_spec(4, width = "10cm")

## Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
## resized.

```

Table 3: Marathon Runners' Data Description

Variables	Missing Data	Type	Description
Age (yr)	0	Numeric	Age (yr) represents the ages of the participants.
age_ranges	NA	Categorical	Age_ranges are the age groups.
Black Globe Temp C	491	Numeric	Black Globe Temp Celcius indicates how hot it feels in direct sunlight. It considers temperature, humidity, wind speed, sun angle, and cloud cover to provide a holistic view of the stress placed on the body in hot environments.

CR	0	HMS/Numeric	CR is the course record for each marathon.
Dew Point in C	491	Numeric	Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form. If the air were to be cooled even more, water vapor would have to come out of the atmosphere in the liquid form, usually as fog or precipitation.
Dry bulb Temp C	491	Numeric	Dry bulb Temp Celcius is the air temperature without taking in account of the humidity or any moisture.
Flag	491	Character	Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT > 10-18C, Yellow=WBGT >18-23C, Red= WBGT >23-28C, and Black= WBGT > 28C
Gender	0	Character	Gender is represented by F= Female and M= Male.
Percent CR	0	Numeric	Percent CR is the percent off current course record by gender.
Percent Relative Humidity	491	Numeric	Percent Relative Humidity how much moisture is in the air compared to the maximum amount of moisture the air can hold at a given temperature. Gives an idea of how humid it feels outside.
Race	0	Character	Race represents the marathons the participants competed, including the B=Boston Marathon, C= Chicago Marathon, NY= New York City Marathon,T= Twin Cities Marathon (Minneapolis,MN), D= Grandma's Marathon (Duluth, MN).
Race_Seconds	0	Numeric	Race_Seconds is the course record measured in seconds.
Runtimes	0	Numeric	Runtimes is the converted gender percentage into seconds. The represent the marathon runners runtime in seconds.
Solar Radiation	491	Numeric	Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels through space and reaches the Earth as light and heat.
Wet Bulb Globe Temp	491	Numeric	Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar radiation on humans. Formula WBGT= 0.7 x Tw + 0.2 x Tg+ 0.1 xTd.
Wet bulb Temp C	491	Numeric	Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity in the air. Wet bulb temperature gives you an idea of how temperature feels when you take humidity into account.
Wind Speed	491	Numeric	Wind Speed in Km/hr.
Year	0	Numeric	Years represented in the dataset ranging from 1993-2016.

```
#Remove all the Na's from the dataset
course_record_project1<- na.omit(course_record_project1)
```

#AIM 1: Examine effects of increasing age on marathon performance in men and women AIM 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

In the course record data we have Percent CR which give percentages by gender In the project 1 data we have course record overall has race seconds ## Aim 1: Examine effects of increasing age on marathon performance in men and women.

```
#Find the minimum age
min(course_record_project1$`Age (yr)`)
```

```
## [1] 14
```

Table 2: Marathon Runners' Data Description

Variables	Missing Data	Type	Description
Age (yr)	0	Numeric	Age (yr) represents the ages of the participants.
age_ranges	NA	Categorical	Age_ranges are the age groups.
Black Globe Temp C	491	Numeric	Black Globe Temp Celcius indicates how hot it feels in direct sunlight.It considers temperature, humidity, wind speed, sun angle, and cloud cover to provide a holistic view of the stress placed on the body in hot environments.
CR	0	HMS/Numeric	CR is the course record for each marathon.
Dew Point in C	491	Numeric	Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form. If the air were to be cooled even more, water vapor would have to come out of the atmosphere in the liquid form, usually as fog or precipitation.
Dry bulb Temp C	491	Numeric	Dry bulb Temp Celcius is the air temperature without taking into account of the humidity or any moisture.
Flag	491	Character	Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C, Red= WBGT >23-28C, and Black= WBGT > 28C
Gender	0	Character	Gender is represented by F= Female and M= Male.
Percent CR	0	Numeric	Percent CR is the percent off current course record by gender.
Percent Relative Humidity	491	Numeric	Percent Relative Humidity how much moisture is in the air compared to the maximum amount of moisture the air can hold at a given temperature. Gives an idea of how humid it feels outside.
Race	0	Character	Race represents the marathons the participants competed, including the B=Boston Marathon, C= Chicago Marathon, NY= New York City Marathon,T= Twin Cities Marathon (Minneapolis,MN), D= Grandma's Marathon (Duluth, MN).
Race_Seconds	0	Numeric	Race_Seconds is the course record measured in seconds.
Runtimes	0	Numeric	Runtimes is the converted gender percentage into seconds. This represent the marathon runners runtime in seconds.
Solar Radiation	491	Numeric	Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels through space and reaches the Earth as light and heat.
Wet Bulb Globe Temp	491	Numeric	Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar radiation on humans. Formula WBGT= $0.7 \times Tw + 0.2 \times Tg + 0.1 \times Td$.
Wet bulb Temp C	491	Numeric	Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity in the air. Wet bulb temperature gives you an idea of how temperature feels when you take humidity into account.
Wind Speed	491	Numeric	Wind Speed in Km/hr.
Year	0	Numeric	¹¹ Years represented in the dataset ranging from 1993-2016.

```

#Find the maximum age
max(course_record_project1$`Age (yr)`)

## [1] 91

#Create Age Ranges/ Age Breaks to Categorize the groups

course_record_project1$age_ranges<- cut(
  course_record_project1$`Age (yr)`,
  breaks = c(0, 25, 35, 45, 55, 65, 75, Inf), # Custom breaks for the new age ranges
  labels = c("<15-25", "26-35", "36-45", "46-55", "56-65", "66-75", "76+")
)

#Get Counts By Age Group to see balance
age_range_counts <- course_record_project1 %>%
  group_by(age_ranges)%>%
  summarise(count=n())

age_range_counts

## # A tibble: 7 x 2
##   age_ranges count
##   <fct>      <int>
## 1 <15-25      1736
## 2 26-35      1840
## 3 36-45      1840
## 4 46-55      1840
## 5 56-65      1820
## 6 66-75      1463
## 7 76+        534

marathon_performance_by_age<- course_record_project1%>%
  select(Race,Year, Gender, age_ranges, Runtimes, `Age (yr)`)%>%
  group_by(Race, Year, Gender, age_ranges)

#Get the best course_record from each race
best_course_race<- course_record_project1 %>%
  filter(Runtimes <= Race_Seconds)%>%
  group_by(Race, Gender, age_ranges)%>%
  summarise(count=n())

## `summarise()` has grouped output by 'Race', 'Gender'. You can override using
## the `.groups` argument.

# Rename some of the column names
best_course_race <- best_course_race %>%
  rename(
    `Marathon` = Race,          # Rename 'Race' to 'Race Name'
    `Gender` = Gender,          # Keep the 'Gender' column as is (optional)
    `Age Range` = age_ranges,   # Rename 'age_ranges' to 'Age Range'
    `Number of Participants` = count # Rename 'count' to 'Number of Participants'
  )

# Create the table with the new column names and specified styling
best_course_race %>%

```

Table 4: `<div style='text-align:center; font-size:24px; font-weight:bold;'>Marathon Runners with High Performance by Race</div>`

Marathon	Gender	Age Range	Number of Participants
B	F	<15-25	2
B	F	26-35	4
B	M	<15-25	2
B	M	26-35	4
C	F	26-35	2
C	M	<15-25	4
C	M	26-35	4
D	F	<15-25	1
D	F	36-45	1
D	M	26-35	1
NY	F	<15-25	1
NY	F	26-35	3
NY	M	<15-25	1
TC	F	26-35	1
TC	F	36-45	2
TC	M	26-35	1

```
kbl(caption = "<div style='text-align:center; font-size:24px; font-weight:bold;'>Marathon Runners with High Performance by Race</div>",
kable_classic(full_width = F, html_font = "Cambria", font_size = 20) %>%
kable_styling(position = "center", font_size = 16))

worst_course_race <- course_record_project1 %>%
  filter(Runtimes >= Race_Seconds) %>%
  group_by(Race, Gender, age_ranges) %>%
  summarise(count=n())

## `summarise()` has grouped output by 'Race', 'Gender'. You can override using
## the `.groups` argument.

just_gender_age_ranges <- course_record_project1 %>%
  filter(Runtimes >= Race_Seconds) %>%
  group_by(Gender, age_ranges) %>%
  summarise(count=n())

## `summarise()` has grouped output by 'Gender'. You can override using the
## `.groups` argument.

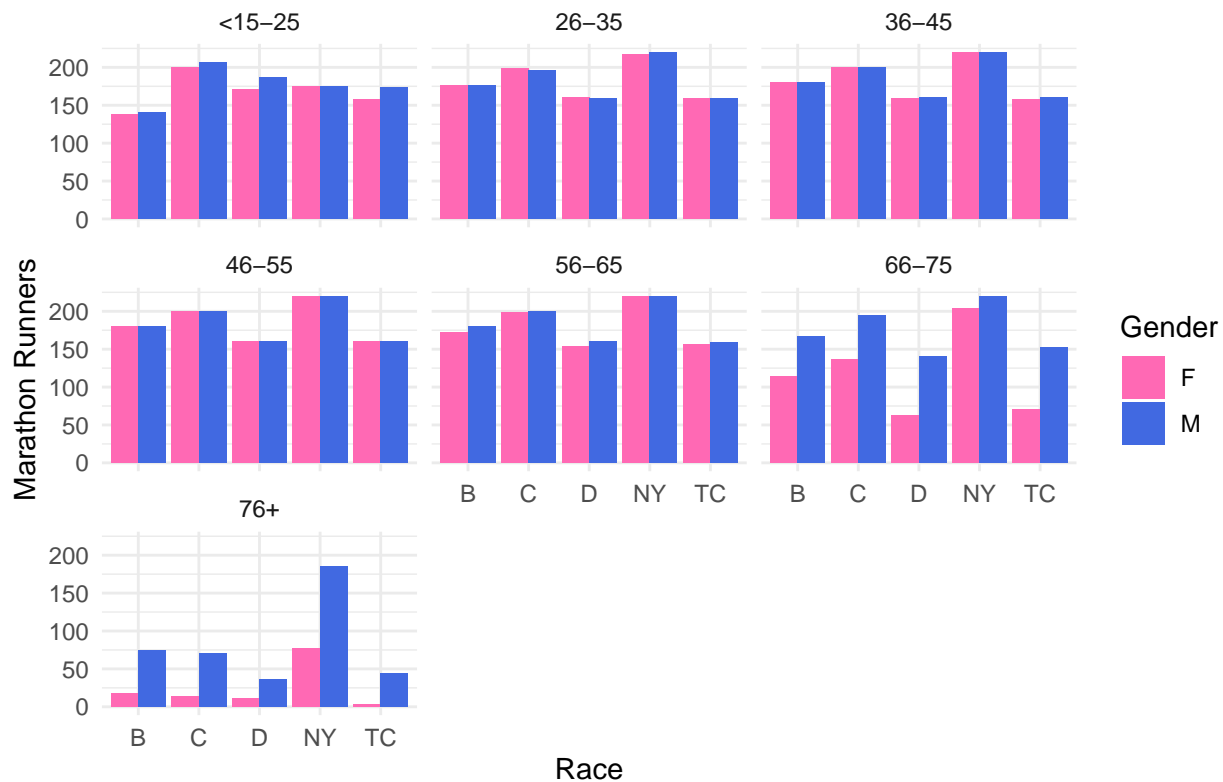
worst_course_race %>%
  kbl(caption = "Number of Marathon Runners that Did not beat the Course Record by Race, Gender, and Age",
kable_classic(full_width = F, html_font = "Times New Roman", font_size = 20))
```

Table 5: Number of Marathon Runners that Did not beat the Course Record by Race, Gender, and Age Group

Race	Gender	age_ranges	count
B	F	<15-25	138
B	F	26-35	176
B	F	36-45	180
B	F	46-55	180
B	F	56-65	172
B	F	66-75	114
B	F	76+	18
B	M	<15-25	141
B	M	26-35	176
B	M	36-45	180
B	M	46-55	180
B	M	56-65	180
B	M	66-75	167
B	M	76+	74
C	F	<15-25	200
C	F	26-35	198
C	F	36-45	200
C	F	46-55	200
C	F	56-65	199
C	F	66-75	137
C	F	76+	14
C	M	<15-25	207
C	M	26-35	196
C	M	36-45	200
C	M	46-55	200
C	M	56-65	200
C	M	66-75	195
C	M	76+	71
D	F	<15-25	171

```
# Create bar plot of the Worst Course Race variable for easier read
ggplot(worst_course_race, aes(x = Race, y = count, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") + # Create a bar plot, grouped by Gender
  labs(title = "Number of Marathon Runners that Did not beat the Course Record by Race, Gender, and Age",
        x = "Race",
        y = "Marathon Runners") +
  scale_fill_manual(values = c("F" = "hotpink", "M" = "royalblue")) + # colors pink for F and blue for M
  facet_wrap(~ age_ranges) +
  theme_minimal()
```

Number of Marathon Runners that Did not beat the Course Record by Race

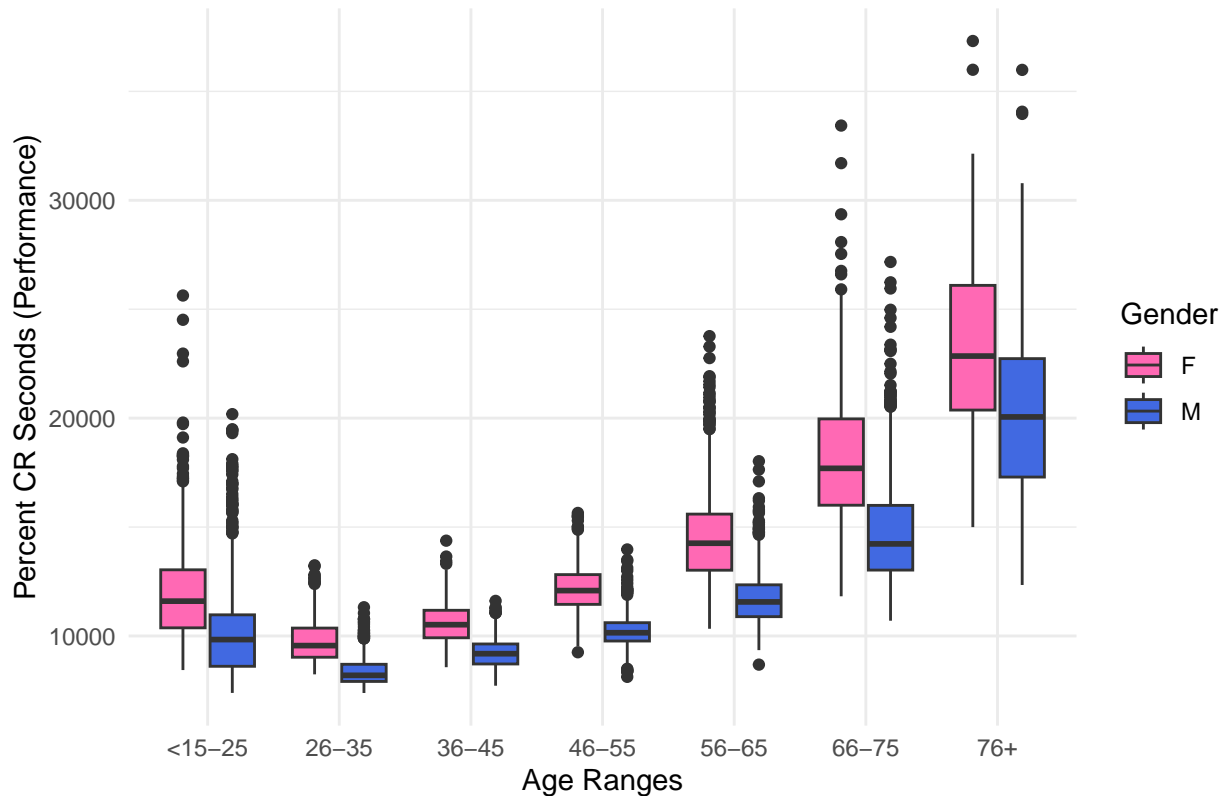


```
library(gridExtra)

# Create the boxplot
age_boxplot<-ggplot(marathon_performance_by_age, aes(x = age_ranges, y =Runtimes, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Effects of Age on Marathon Performance in Men and Women",
        x = "Age Ranges",
        y = "Percent CR Seconds (Performance)") +
  scale_fill_manual(values = c("F" = "hotpink", "M" = "royalblue"))+
  theme_minimal()

age_boxplot
```

Effects of Age on Marathon Performance in Men and Women

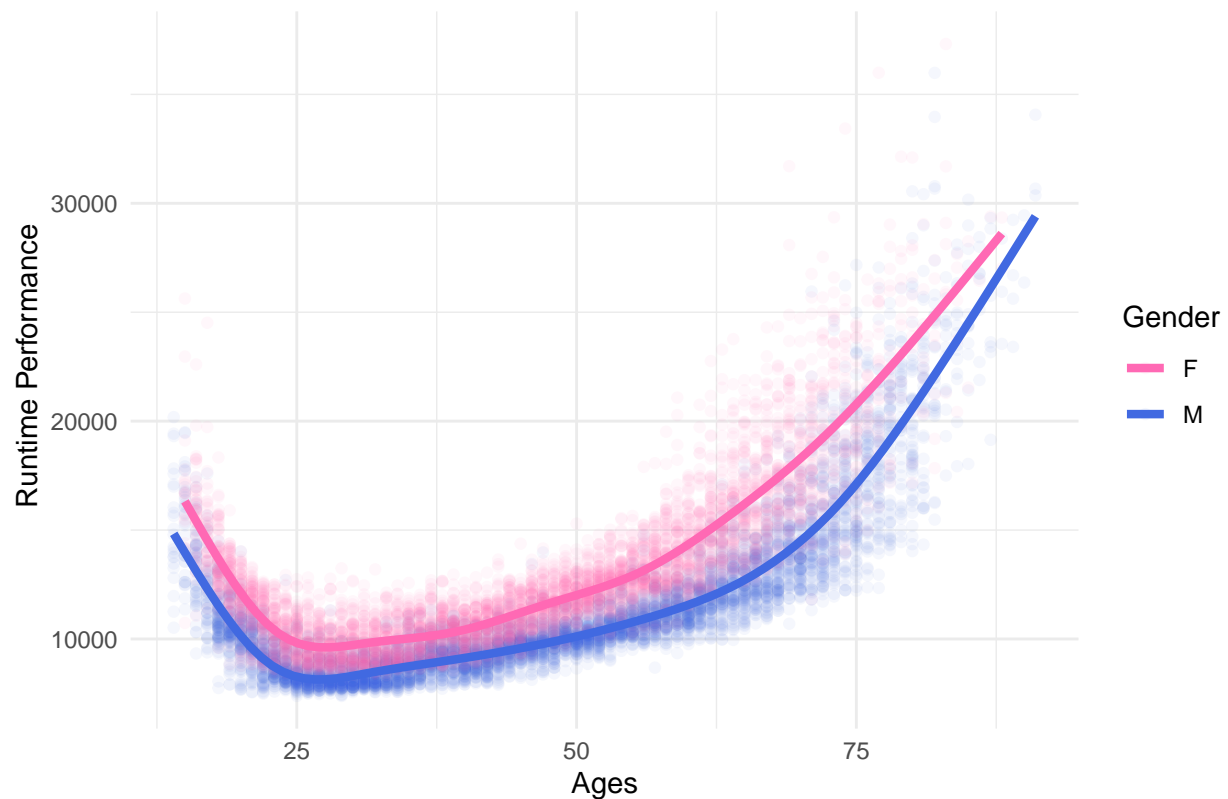


```
age_plot<-ggplot(marathon_performance_by_age, aes(x = `Age (yr)`, y = Runtimes, color = Gender)) +
  geom_point(alpha = 0.05) +
  geom_smooth(se = FALSE, linewidth = 1.5) +
  labs(title = "Effects of Age on Marathon Performance in Men and Women",
        x = "Ages",
        y = " Runtime Performance") +
  scale_color_manual(values = c("F" = "hotpink", "M" = "royalblue")) +
  theme_minimal()
```

```
age_plot
```

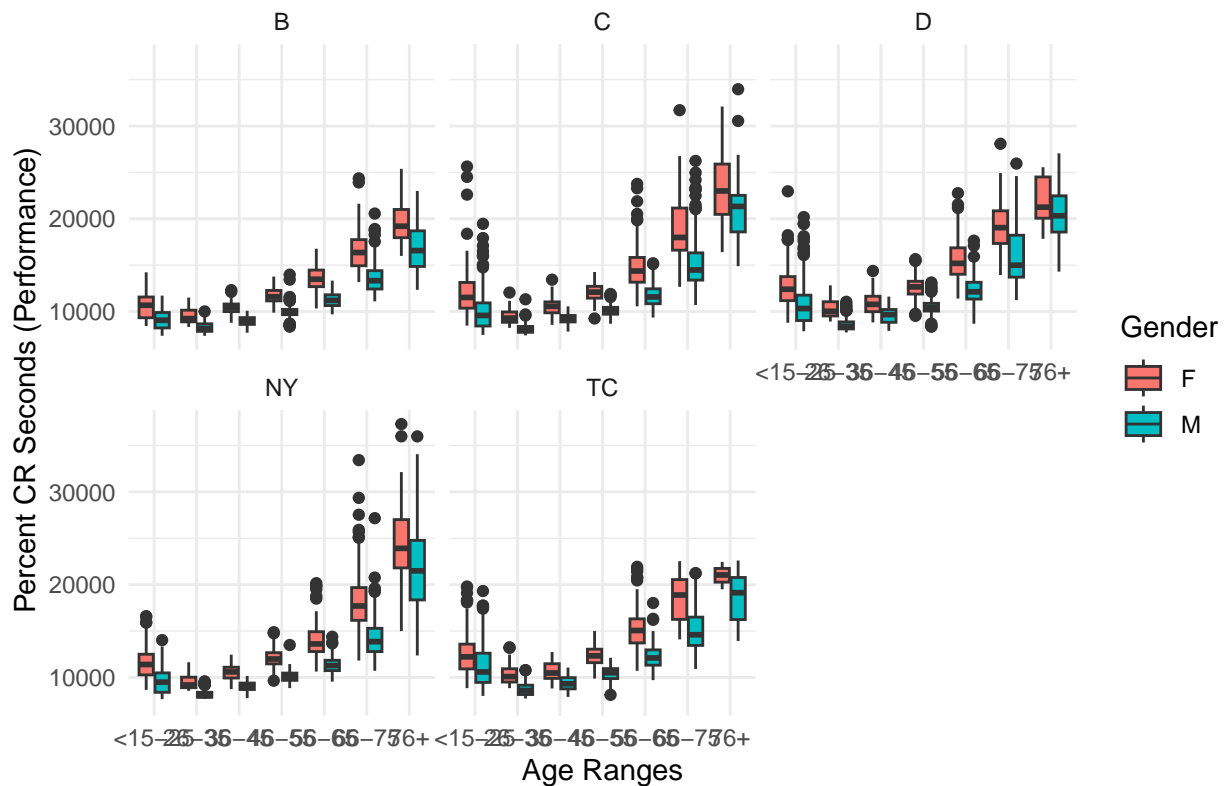
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```


Effects of Age on Marathon Performance in Men and Women



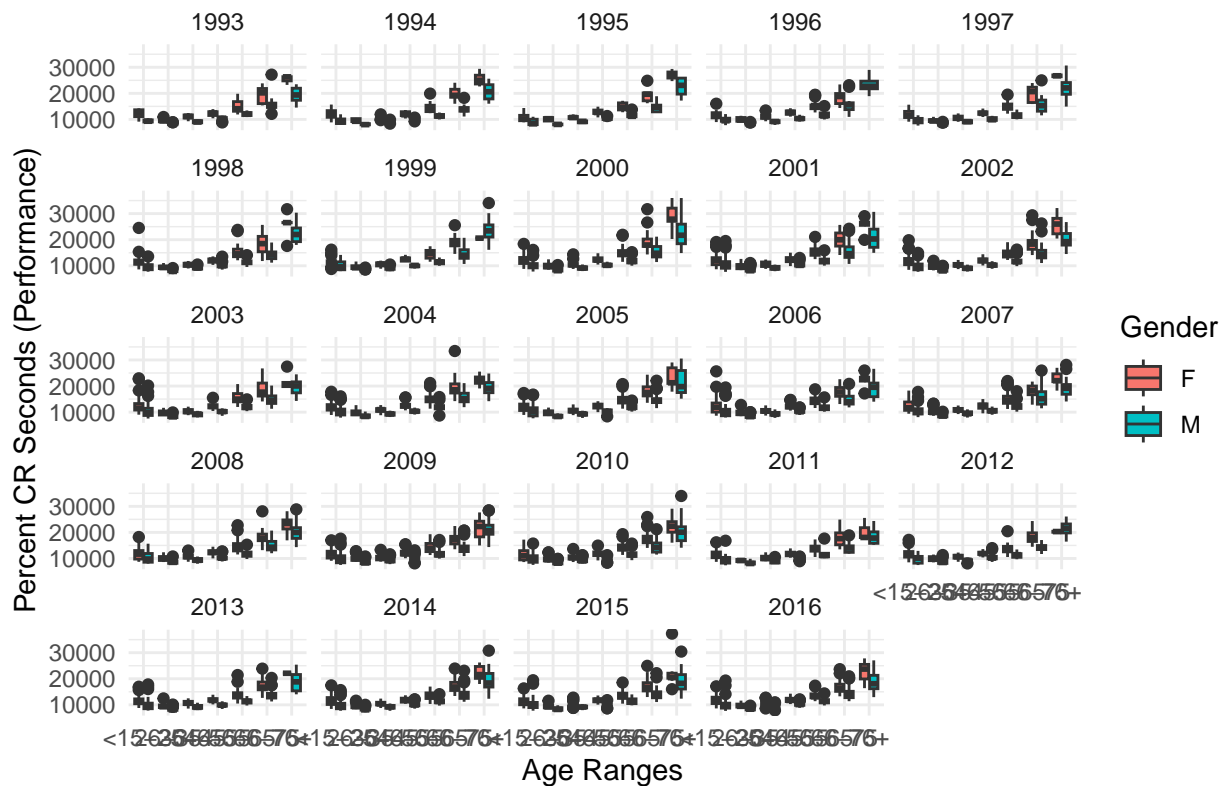
```
# Create the boxplot, faceting by Race
ggplot(marathon_performance_by_age, aes(x = age_ranges, y = Runtimes, fill = Gender)) +
  geom_boxplot() +
  facet_wrap(~ Race) + # Facet by Race to create separate plots for each race
  labs(title = "Effects of Age on Marathon Performance by Race, Gender, and Age",
        x = "Age Ranges",
        y = "Percent CR Seconds (Performance)") +
  theme_minimal()
```

Effects of Age on Marathon Performance by Race, Gender, and Age



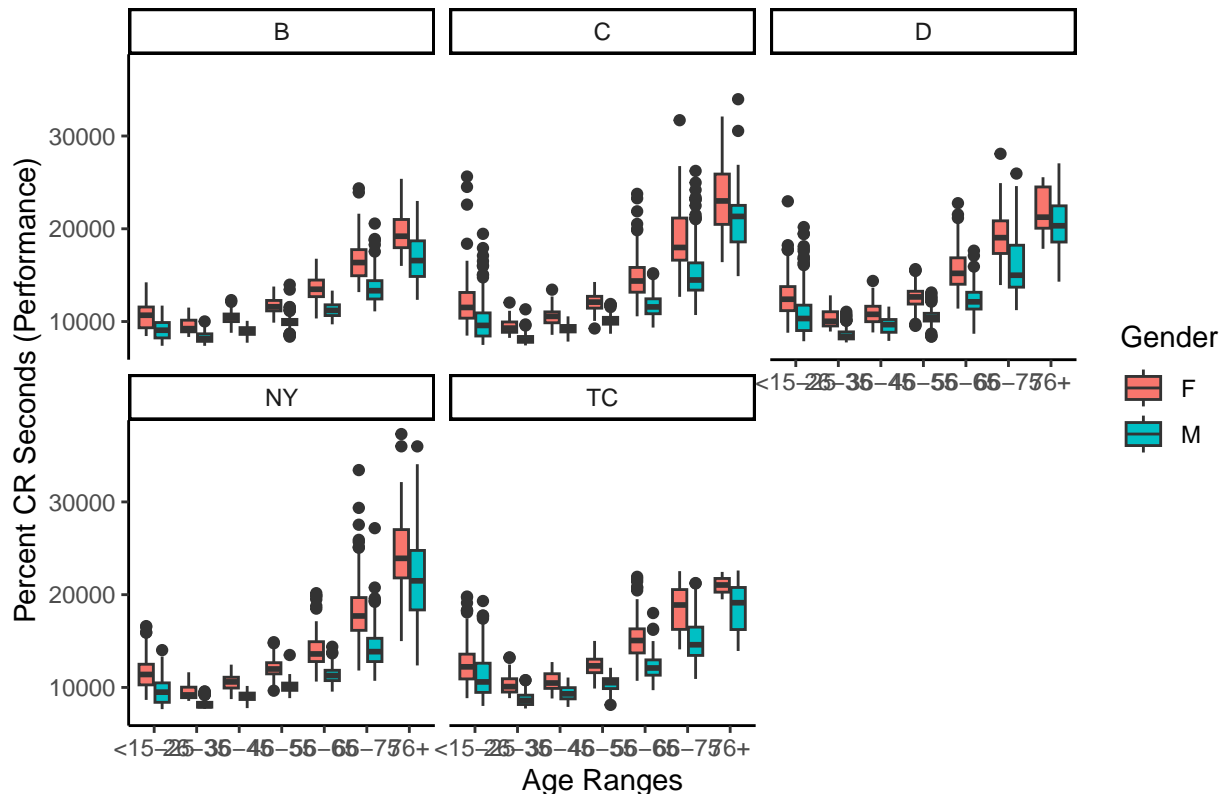
```
# Create the boxplot, faceting by Race
ggplot(marathon_performance_by_age, aes(x = age_ranges, y = Runtimes, fill = Gender)) +
  geom_boxplot() +
  facet_wrap(~ Year) + # Facet by Race to create separate plots for each race
  labs(title = "Effects of Age on Marathon Performance by Race, Gender, and Age by year",
        x = "Age Ranges",
        y = "Percent CR Seconds (Performance)") +
  theme_minimal()
```

Effects of Age on Marathon Performance by Race, Gender, and Age by year



```
# Create the boxplot, faceting by Race
ggplot(marathon_performance_by_age, aes(x = age_ranges, y = Runtimes, fill = Gender)) +
  geom_boxplot() +
  facet_wrap(~ Race) + # Facet by Race to create separate plots for each race
  labs(title = "Effects of Age on Marathon Performance by Race, Gender, and Age",
        x = "Age Ranges",
        y = "Percent CR Seconds (Performance)") +
  theme_classic()
```

Effects of Age on Marathon Performance by Race, Gender, and Age



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

Higher temperatures lead to slower times stratify by gender age

```
environmental_conditions <- course_record_project1 %>%
  select(Race, Gender, `Age (yr)`, Runtimes, `Dry bulb Temp C`, `Wet bulb Temp C`,
    `Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`, `Dew Point in C`,
    `Wind Speed`, `Wet Bulb Globe Temp`, age_ranges) %>%
  group_by( Gender, `Age (yr)`)
```

#runs stuff that doesnt need to run for knitting purposes

Linear Model Approach

```
# # Ensure Gender is a factor if not already
# environmental_conditions$Gender <- as.factor(environmental_conditions$Gender)
#
```

Linear to evaluate statistical significance

```
model <- lm(Runtimes ~ `Dry bulb Temp C` + `Percent Relative Humidity` + `Black Globe Temp C` +
  `Wind Speed` + `Dew Point in C` + `Solar Radiation` + `Age (yr)` + `Wet Bulb Globe Temp` + Gender +
  `Dry bulb Temp C`:`Age (yr)` + `Percent Relative Humidity`:`Age (yr)` +
  `Black Globe Temp C`:`Age (yr)` + `Wind Speed`:`Age (yr)` + `Dew Point in C`:`Age (yr)` +
  `Solar Radiation`:`Age (yr)` + `Wet Bulb Globe Temp`:`Age (yr)` + Gender:`Age (yr)` +
  `Dry bulb Temp C`:Gender + `Percent Relative Humidity`:Gender + `Black Globe Temp C`:Gender +
  `Wind Speed`:Gender + `Dew Point in C`:Gender + `Solar Radiation`:Gender +
  `Wet Bulb Globe Temp`:Gender, data = environmental_conditions)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Runtimes ~ `Dry bulb Temp C` + `Percent Relative Humidity` +
##   `Black Globe Temp C` + `Wind Speed` + `Dew Point in C` +
##   `Solar Radiation` + `Age (yr)` + `Wet Bulb Globe Temp` +
##   Gender + `Dry bulb Temp C`:`Age (yr)` + `Percent Relative Humidity`:`Age (yr)` +
##   `Black Globe Temp C`:`Age (yr)` + `Wind Speed`:`Age (yr)` +
##   `Dew Point in C`:`Age (yr)` + `Solar Radiation`:`Age (yr)` +
##   `Wet Bulb Globe Temp`:`Age (yr)` + Gender:`Age (yr)` + `Dry bulb Temp C`:Gender +
##   `Percent Relative Humidity`:Gender + `Black Globe Temp C`:Gender +
##   `Wind Speed`:Gender + `Dew Point in C`:Gender + `Solar Radiation`:Gender +
##   `Wet Bulb Globe Temp`:Gender, data = environmental_conditions)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5705.5 -1570.9  -779.3   661.6 19133.0
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.412e+03  5.787e+02   4.167 3.10e-05
## `Dry bulb Temp C`              -5.529e+02  1.228e+02  -4.504 6.74e-06
## `Percent Relative Humidity`    2.107e+01  2.970e+00   7.092 1.40e-12
## `Black Globe Temp C`          -2.705e+02  6.508e+01  -4.157 3.26e-05
## `Wind Speed`                  -2.468e+01  2.122e+01  -1.163 0.244853
## `Dew Point in C`              -3.136e+02  9.241e+01  -3.394 0.000692
## `Solar Radiation`             4.711e+00  5.778e-01   8.153 3.95e-16
## `Age (yr)`                    2.215e+02  1.116e+01  19.839 < 2e-16
## `Wet Bulb Globe Temp`         1.289e+03  2.904e+02   4.438 9.15e-06
## GenderM                       -1.891e+03  4.191e+02  -4.512 6.49e-06
## `Dry bulb Temp C`:`Age (yr)`   1.040e+01  2.406e+00   4.323 1.55e-05
## `Percent Relative Humidity`:`Age (yr)` -5.532e-01  5.694e-02  -9.714 < 2e-16
## `Black Globe Temp C`:`Age (yr)`  4.724e+00  1.283e+00   3.681 0.000233
## `Wind Speed`:`Age (yr)`        3.165e-01  4.085e-01   0.775 0.438447
## `Dew Point in C`:`Age (yr)`    4.888e+00  1.814e+00   2.695 0.007055
## `Solar Radiation`:`Age (yr)`   -1.242e-01  1.132e-02 -10.974 < 2e-16
## `Age (yr)`:`Wet Bulb Globe Temp` -2.146e+01  5.701e+00  -3.765 0.000168
## `Age (yr)`:GenderM            -6.546e+00  2.735e+00  -2.393 0.016711
## `Dry bulb Temp C`:GenderM      -3.656e+01  8.562e+01  -0.427 0.669373
## `Percent Relative Humidity`:GenderM  4.822e+00  2.049e+00   2.353 0.018618
## `Black Globe Temp C`:GenderM   -2.462e+01  4.515e+01  -0.545 0.585528
## `Wind Speed`:GenderM           7.716e+00  1.453e+01   0.531 0.595476
## `Dew Point in C`:GenderM      -3.339e+01  6.405e+01  -0.521 0.602218
## `Solar Radiation`:GenderM      6.530e-01  3.992e-01   1.636 0.101958
## `Wet Bulb Globe Temp`:GenderM   8.433e+01  2.020e+02   0.417 0.676408
##
## (Intercept) ***
## `Dry bulb Temp C` ***
## `Percent Relative Humidity` ***
## `Black Globe Temp C` ***
## `Wind Speed`
## `Dew Point in C` ***
```

```

## `Solar Radiation` ***
## `Age (yr)` ***
## `Wet Bulb Globe Temp` ***
## GenderM ***
## `Dry bulb Temp C`: `Age (yr)` ***
## `Percent Relative Humidity`: `Age (yr)` ***
## `Black Globe Temp C`: `Age (yr)` ***
## `Wind Speed`: `Age (yr)`
## `Dew Point in C`: `Age (yr)` **
## `Solar Radiation`: `Age (yr)` ***
## `Age (yr)`: `Wet Bulb Globe Temp` ***
## `Age (yr)`: GenderM *
## `Dry bulb Temp C`: GenderM
## `Percent Relative Humidity`: GenderM *
## `Black Globe Temp C`: GenderM
## `Wind Speed`: GenderM
## `Dew Point in C`: GenderM
## `Solar Radiation`: GenderM
## `Wet Bulb Globe Temp`: GenderM
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2536 on 11048 degrees of freedom
## Multiple R-squared:  0.5218, Adjusted R-squared:  0.5207
## F-statistic: 502.3 on 24 and 11048 DF,  p-value: < 2.2e-16

# Load necessary libraries
library(ggplot2)
library(corrplot) # For creating a correlation plot

## corrplot 0.94 loaded

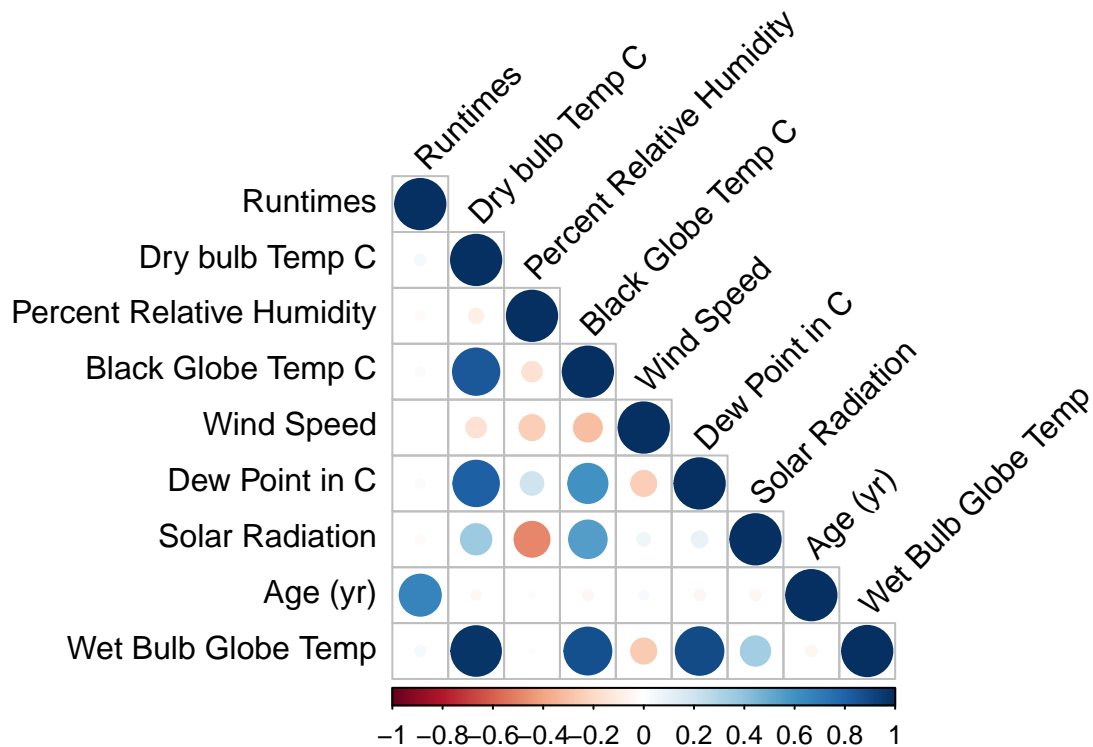
# Assuming course_records is your data frame

# Step 1: Select only numeric columns from the data frame
numeric_data <- course_record_project1%>%
  select(Runtimes, `Dry bulb Temp C`, `Percent Relative Humidity`, `Black Globe Temp C`,
    `Wind Speed`, `Dew Point in C`, `Solar Radiation`, `Age (yr)`, `Wet Bulb Globe Temp`)

# Step 2: Compute the correlation matrix
cor_matrix <- cor(numeric_data, use = "complete.obs") # Use complete.obs to ignore NAs

# Step 3: Visualize the correlation matrix using corrplot
corrplot(cor_matrix, method = "circle", type = "lower", tl.col = "black", tl.srt = 45)

```

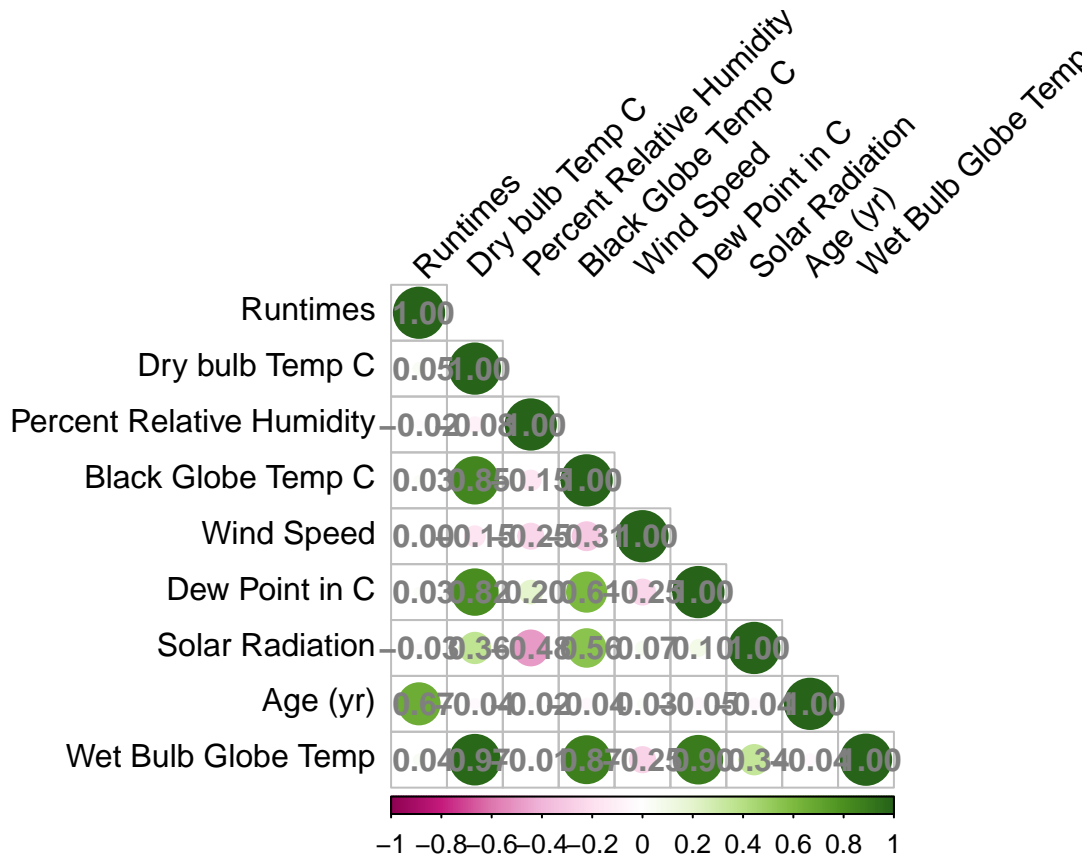


```
# Melt the correlation matrix for ggplot2
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
## The following object is masked from 'package:tidyr':
##
## smiths
```

```
cor_data <- melt(cor_matrix)
```

```
# Create the ggplot correlation heatmap
# Break down and Visualization of the correlation matrix with environmental on age variable names
corrplot(cor_matrix,
  method = "circle",           # Use circle method for visualization
  type = "lower",              # Only display lower triangle
  tl.col = "black",            # Color of the text labels
  tl.srt = 45,                 # Rotate the text labels at 45 degrees
  tl.pos = "lt",               # Display text labels on left and top
  col = COL2('PiYG'),          # Color palette for the plot
  cl.pos = 'b',                # Place the color legend at the bottom
  addCoef.col = 'grey50',      # Add correlation coefficient values in grey
  is.corr = TRUE,              # Ensure this is treated as a correlation matrix
  col.lim = c(-1, 1)          # Set the color limit for correlation values
)
```

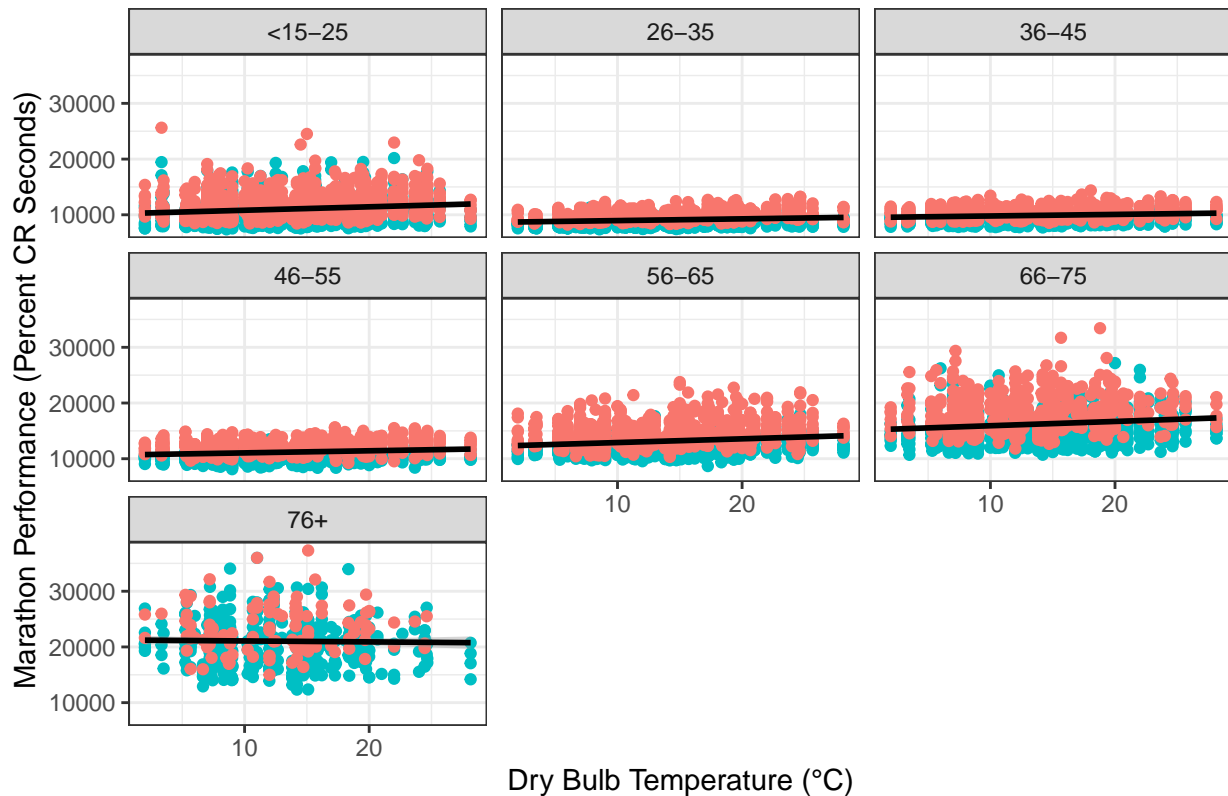


```
# Dry bulb plot
dry_bulb_plot<-ggplot(enviromental_conditions, aes(x = `Dry bulb Temp C`, y = Runtimes, color = Gender)) +
  geom_point() + # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confidence interval
  # Adding labels with gender
  facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
  theme_bw() + # Black and white theme
  guides(color = 'none') + # Remove legend for color
  labs(title = "Marathon Performance vs. Dry Bulb Temperature by Age Ranges",
        x = "Dry Bulb Temperature (°C)",
        y = "Marathon Performance (Percent CR Seconds)") +
  scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))

dry_bulb_plot
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's fill values.
```


Marathon Performance vs. Dry Bulb Temperature by Age Ranges

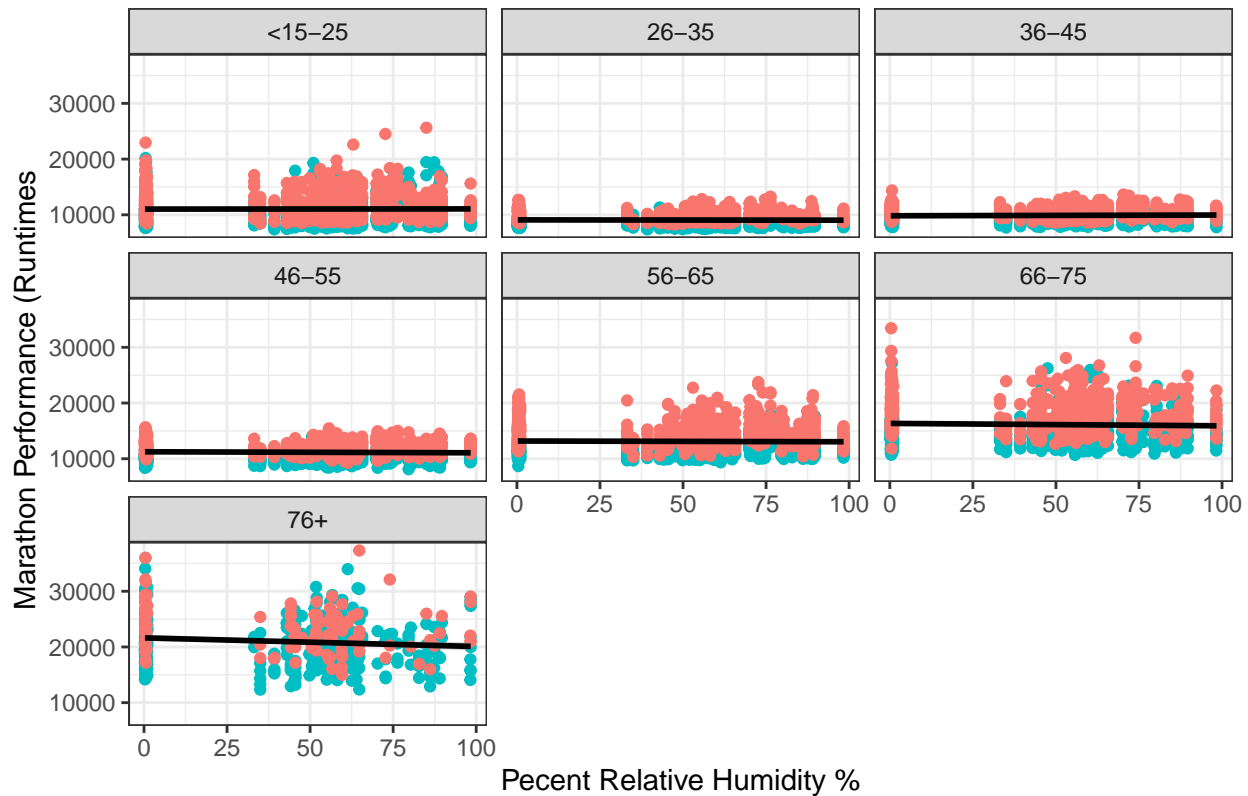


```
# Humidity
humidity_runtimes<- ggplot(enviromental_conditions, aes(x = `Percent Relative Humidity`, y = Runtimes,
  geom_point() + # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confide
# Adding labels with gender
  facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
  theme_bw() + # Black and white theme
  guides(color = 'none') + # Remove legend for color
  labs(title = "Marathon Performance vs. Humidity Stratified by Age Ranges ",
    x = "Pecent Relative Humidity %",
    y = "Marathon Performance (Runtimes)")+
  scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))

humidity_runtimes
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's fill values.
```

Marathon Performance vs. Humidity Stratified by Age Ranges

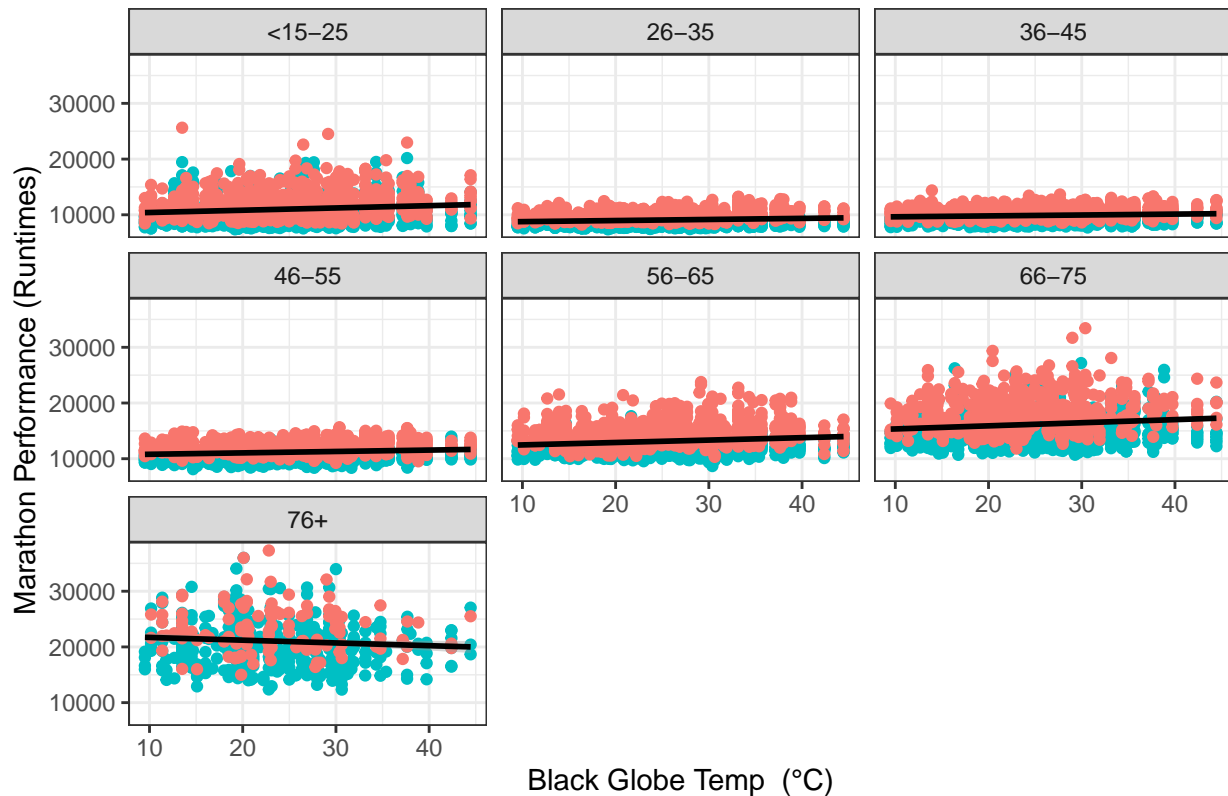


```
# Black Globe Temperature
black_globe_temp_graph<- ggplot(enviromental_conditions, aes(x = `Black Globe Temp C`, y = Runtimes, color = gender)) +
  geom_point() + # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confidence interval
  # Adding labels with gender
  facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
  theme_bw() + # Black and white theme
  guides(color = 'none') + # Remove legend for color
  labs(title = "Marathon Performance vs. Dry Bulb Temperature Stratified by Age Groups",
        x = "Black Globe Temp (°C)",
        y = "Marathon Performance (Runtimes)") +
  scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))

black_globe_temp_graph
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's fill values.
```

Marathon Performance vs. Dry Bulb Temperature Stratified by Age Group

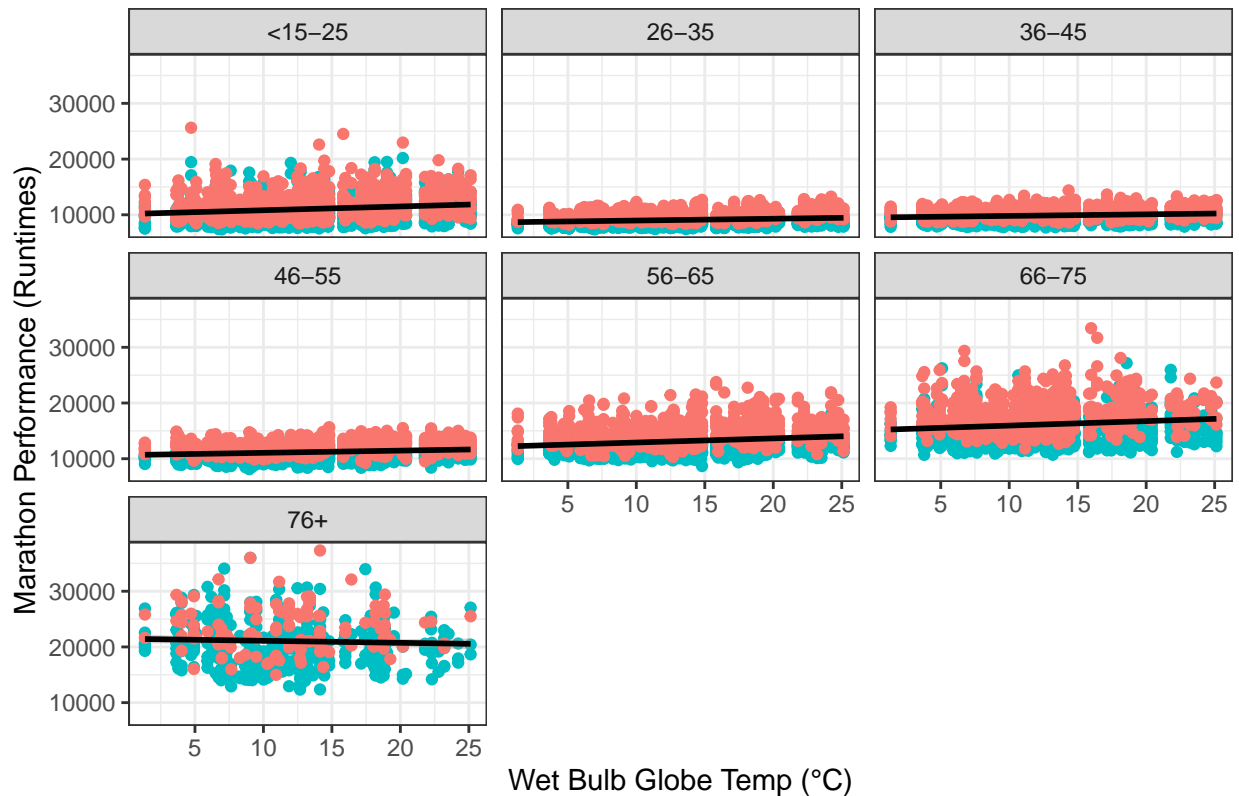


```
# Wet Bulb Temperature
wet_bulb_graph<-ggplot(environmental_conditions, aes(x = `Wet Bulb Globe Temp`,
                                                    y = Runtimes, color = Gender)) +geom_point() + # Scatter plot
                                                    geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regress
# Adding labels with gender
facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
theme_bw() + # Black and white theme
guides(color = 'none') + # Remove legend for color
labs(title = "Marathon Performance vs. Wet Bulb Globe Temp Stratified by Age Groups",
     x = "Wet Bulb Globe Temp (°C)",
     y = "Marathon Performance (Runtimes)")+
scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))

wet_bulb_graph
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's fill values.
```

Marathon Performance vs. Wet Bulb Globe Temp Stratified by Age Group



```
#Solar Radiation Graph
solar_radiation_graph<-ggplot(environmental_conditions, aes(x = `Solar Radiation`,
  y = Runtimes, color = Gender)) +geom_point() + # Scatter plot
  geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regress

# Adding labels with gender
facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
theme_bw() + # Black and white theme
guides(color = 'none') + # Remove legend for color
labs(title = "Marathon Performance vs. Solar Radiation Stratified by Age Groups",
  x = "Solar Radiation",
  y = "Marathon Performance (Runtimes)")+
scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))
```

```
head(marathon_dates)
```

```
## # A tibble: 6 x 3
##   marathon date      year
##   <chr>      <date>    <dbl>
## 1 Boston  1998-04-20  1998
## 2 Boston  1999-04-19  1999
## 3 Boston  2000-04-17  2000
## 4 Boston  2001-04-16  2001
## 5 Boston  2002-04-15  2002
## 6 Boston  2003-04-21  2003
```

```
#Change Race Names
marathon_dates$Race
```

```

## Warning: Unknown or uninitialised column: `Race`.
## NULL
marathon_dates <- marathon_dates %>%
  mutate(marathon = case_when(
    marathon == "NYC" ~ "NY",
    marathon == "Grandmas" ~ "D",
    marathon == "Boston" ~ "B",
    marathon == "Twin Cities" ~ "TC"
  ))

colnames(marathon_dates)[colnames(marathon_dates) == "marathon"] = "Race"
colnames(marathon_dates)[colnames(marathon_dates) == "year"] = "Year"

# Change formatting of the dates
marathon_dates <- marathon_dates %>%
  mutate(date = as.Date(date, format = "%Y-%m-%d"))

# Combine the marathon dates dataframe to my current dataframe by using left_join
course_record_project1 <- course_record_project1 %>%
  left_join(marathon_dates, by = c("Race", "Year"))

names(aqi_values)

## [1] "cbsa_code"      "state_code"      "county_code"      "site_number"
## [5] "date_local"     "parameter_code"   "units_of_measure" "sample_duration"
## [9] "aqi"            "arithmetic_mean" "marathon"
aqi_values

## # A tibble: 10,451 x 11
##   cbsa_code state_code county_code site_number date_local parameter_code
##   <dbl>      <dbl> <chr>      <chr>      <date>      <dbl>
## 1    14460         25 017         1801    1998-04-20    44201
## 2    14460         25 017         1801    1998-04-20    44201
## 3    14460         25 017         1801    1998-04-20    44201
## 4    14460         25 017         1801    1998-04-20    44201
## 5    14460         25 009         0005    1998-04-20    44201
## 6    14460         25 025         1003    1998-04-20    44201
## 7    14460         25 009         2006    1998-04-20    44201
## 8    14460         25 009         4004    1998-04-20    44201
## 9    14460         25 017         1102    1998-04-20    44201
## 10   14460         25 017         4003    1998-04-20    44201
## # i 10,441 more rows
## # i 5 more variables: units_of_measure <chr>, sample_duration <chr>, aqi <dbl>,
## #   arithmetic_mean <dbl>, marathon <chr>

# Change the marathon variable to Race to match corresponding data and change race names
aqi_values <- aqi_values %>%
  rename(Race = marathon) %>%
  mutate(
    Race = case_when(
      Race == "NYC" ~ "NY",
      Race == "Grandmas" ~ "D",
      Race == "Boston" ~ "B",

```

```

    Race == "Twin Cities" ~"TC"
  ),
  date = as.Date(date_local, format = "%Y-%m-%d"),
  Year = as.numeric(format(date, "%Y"))
) %>%
select(-date_local) #Remove the date_local variable

# calculate average ozone ppm (8-hour avg)
avg_ppm <- aqi_values %>%
  filter(units_of_measure == "Parts per million",
    sample_duration == "8-HR RUN AVG BEGIN HOUR") %>%
  group_by(Race, Year, date) %>%
  summarize(avg_ppm = mean(arithmetic_mean, na.rm = T)) %>%
  ungroup()

## `summarise()` has grouped output by 'Race', 'Year'. You can override using the
## `.groups` argument.

# Merge data_frame to current dataframe
course_record_project1 <- course_record_project1 %>%
  left_join(avg_ppm, by = c("Race", "Year", "date"))

```

AIM3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

```

weather_parameters <- course_record_project1 %>%
  select(Gender, `Age (yr)`, Runtimes, `Dry bulb Temp C`, `Wet bulb Temp C`,
    `Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`, `Dew Point in C`,
    `Wind Speed`, `Wet Bulb Globe Temp`, age_ranges, Flag)

#
# flag_cat <- ggplot(weather_parameters, aes(x = Flag, y = Runtimes, color = Gender)) +
#   boxplot() + # Scatter plot
#   #geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with conf
#   # Adding labels with gender
#   #facet_wrap(~ age_ranges) + # Facet the plot by age_ranges
#   theme_bw() + # Black and white theme
#   guides(color = 'none') + # Remove legend for color
#   labs(title = "Marathon Performance vs. Dry Bulb Temperature Stratified by Age Groups",
#     x = "Black Globe Temp (°C)",
#     y = "Marathon Performance (Runtimes)")

# Filter and prepare the data
weather_parameters <- course_record_project1 %>%
  select(Gender, `Age (yr)`, Runtimes, `Dry bulb Temp C`, `Wet bulb Temp C`,
    `Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`,
    `Dew Point in C`, `Wind Speed`, `Wet Bulb Globe Temp`, age_ranges, Flag)

```

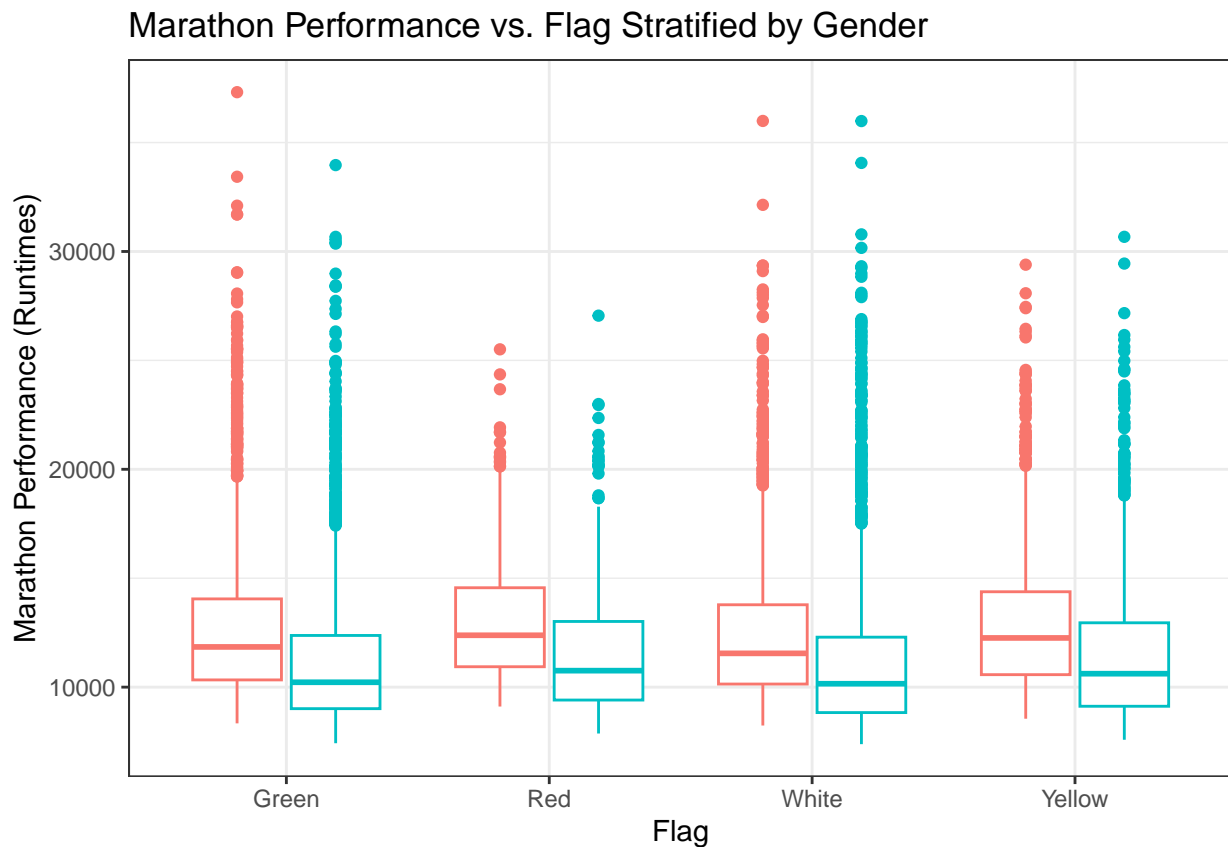
```

# Create the box plot
flag_cat <- ggplot(weather_parameters, aes(x = Flag, y = Runtimes, color = Gender)) +
  geom_boxplot() + # Use geom_boxplot for creating a boxplot
  theme_bw() +     # Black and white theme for a clean plot
  guides(color = 'none') + # Remove legend for color (if not needed)
  labs(
    title = "Marathon Performance vs. Flag Stratified by Gender",
    x = "Flag",
    y = "Marathon Performance (Runtimes)"
  )

# If you want to stratify by age_ranges, uncomment the following line:
# + facet_wrap(~ age_ranges)

# Plot the result
print(flag_cat)

```



```
as.factor(course_record_project1$Flag)
```

```

##      [1] Green  Green  Green  Green  Green  Green  Green  Green  Green  Green  Green
##      [11] Green  Green  Green  Green  Green  Green  Green  Green  Green  Green  Green
##      [21] Green  Green  Green  Green  Green  Green  Green  Green  Green  Green  Green
##      [31] Green  Green  Green  Green  Green  Green  Green  Green  Green  Green  Green
##      [41] Green  Green  Green  Green  Green  Green  Green  Green  Green  Green  Green
##      [51] Green  Green  Green  Green  Green  Green  Green  Green  Green  Green  Green
##      [61] Green  Green  Green  Green  Green  White  White  White  White  White  White
##      [71] White  White  White  White  White  White  White  White  White  White  White

```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

## [10881] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [10891] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [10901] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [10911] Yellow Yellow Yellow Yellow Yellow Green Green Green Green Green
## [10921] Green Green Green Green Green Green Green Green Green Green Green
## [10931] Green Green Green Green Green Green Green Green Green Green Green
## [10941] Green Green Green Green Green Green Green Green Green Green Green
## [10951] Green Green Green Green Green Green Green Green Green Green Green
## [10961] Green Green Green Green Green Green Green Yellow Yellow Yellow
## [10971] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [10981] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [10991] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [11001] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [11011] Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow Yellow
## [11021] Green Green Green Green Green Green Green Green Green Green Green
## [11031] Green Green Green Green Green Green Green Green Green Green Green
## [11041] Green Green Green Green Green Green Green Green Green Green Green
## [11051] Green Green Green Green Green Green Green Green Green Green Green
## [11061] Green Green Green Green Green Green Green Green Green Green Green
## [11071] Green Green Green
## Levels: Green Red White Yellow

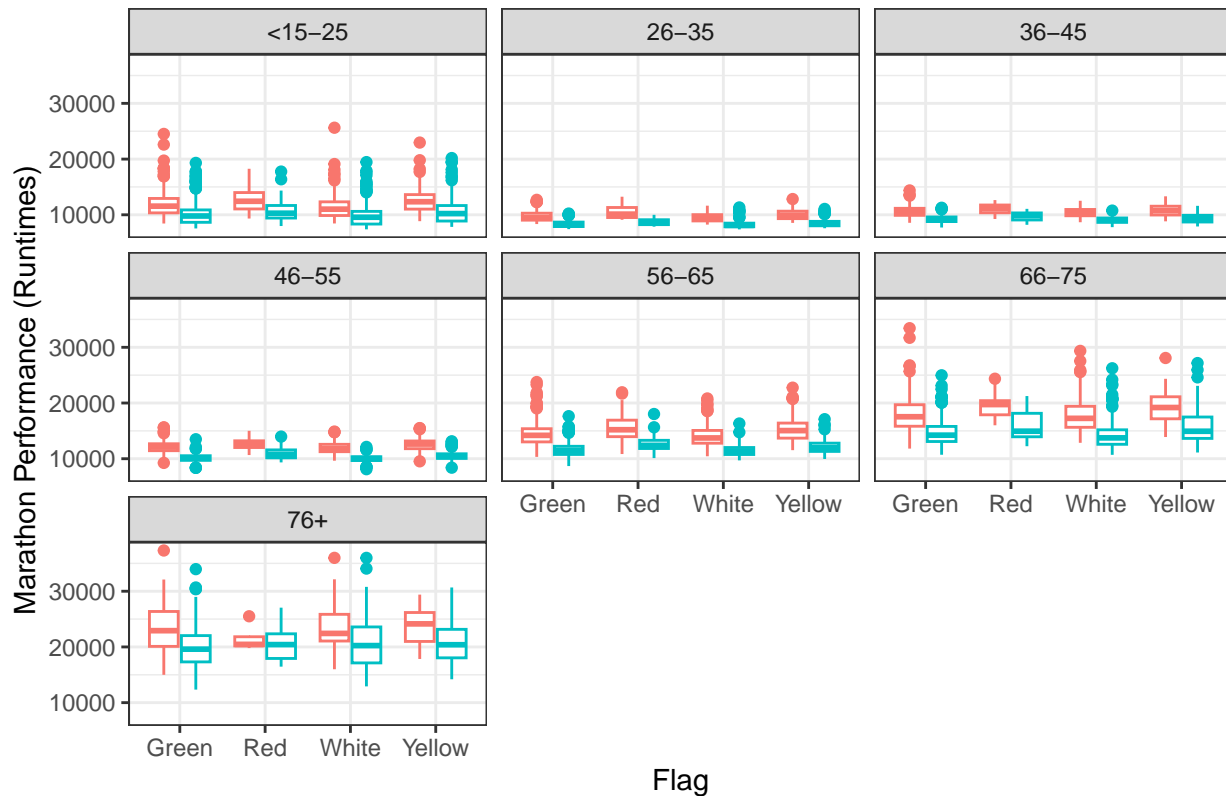
# Create the box plot
flag_cat_age_ranges <- ggplot(weather_parameters, aes(x = Flag, y = Runtimes, color = Gender)) +
  geom_boxplot() +
  facet_wrap(~ age_ranges) + # Use geom_boxplot for creating a boxplot
  theme_bw() + # Black and white theme for a clean plot
  guides(color = 'none') + # Remove legend for color (if not needed)
  labs(
    title = "Marathon Performance vs. Flag Stratified by Gender",
    x = "Flag",
    y = "Marathon Performance (Runtimes)"
  )

# If you want to stratify by age_ranges, uncomment the following line:
# + facet_wrap(~ age_ranges)

# Plot the result
print(flag_cat_age_ranges)

```

Marathon Performance vs. Flag Stratified by Gender



```
#
# flag_exam <- ggplot(weather_parameters, aes(x = Gender, y = Runtimes, color = Flag)) +
#   geom_boxplot() + # Use geom_boxplot for creating a boxplot
#   theme_bw() +    # Black and white theme for a clean plot
#   guides(color = 'none') + # Remove legend for color (if not needed)+
#   scale_color_manual(values = c(
#     "White" = "white", # Replace these with the Flag categories and the colors you want to assign
#     "Green" = "green",
#     "Black" = "black",
#     "Yellow" = "yellow", # Add more colors for other Flag categories as needed
#     "Red" = "red"
#   )) +
#   labs(
#     title = "Marathon Performance by Flag",
#     x = "Flag",
#     y = "Marathon Performance (Runtimes)",
#     color = "Flag Colors"+
#     fill = "Flag Color" # Title for the legend
#   ) +
#   theme(legend.position = "right")

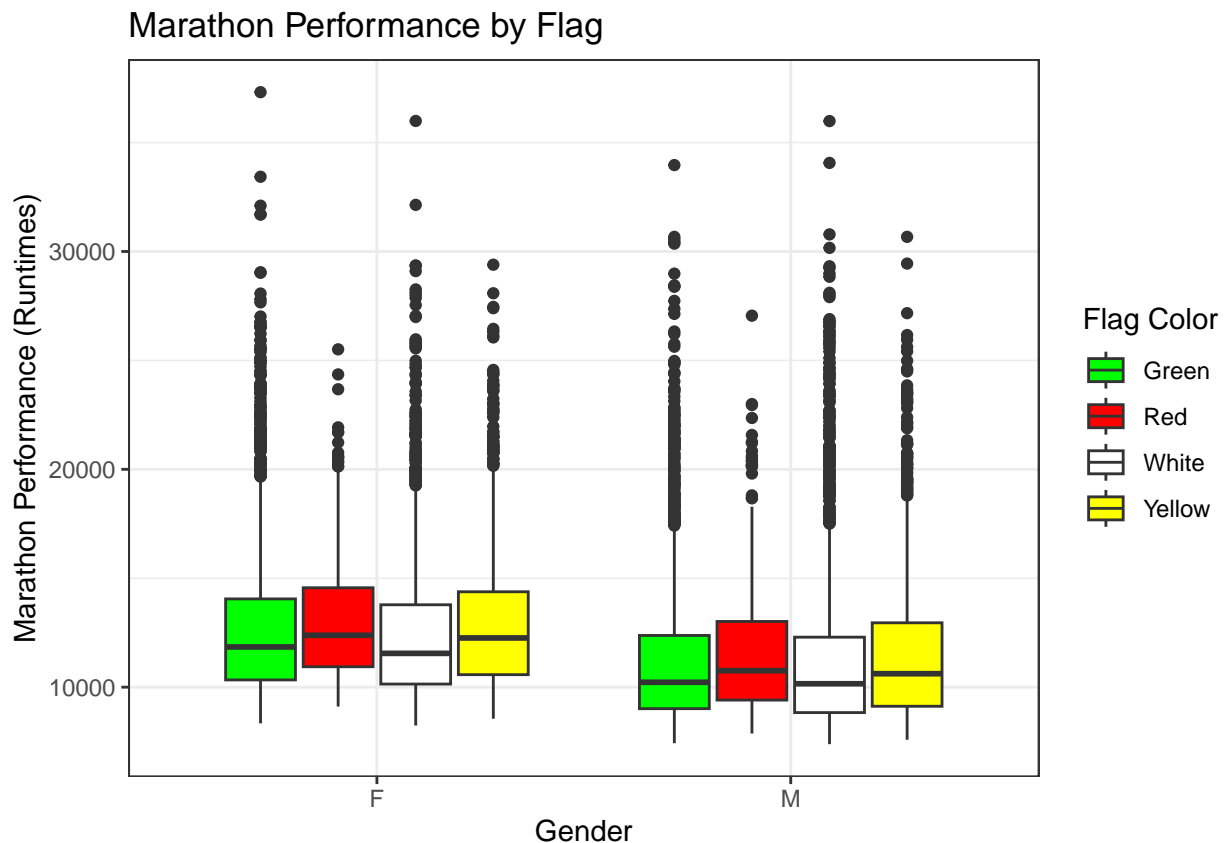
# Create the box plot with custom Flag colors and filled boxplots
flag_exam <- ggplot(weather_parameters, aes(x = Gender, y = Runtimes, fill = Flag)) +
  geom_boxplot() + # Use geom_boxplot for creating a boxplot
  theme_bw() +    # Black and white theme for a clean plot
  scale_fill_manual(values = c(
```

```

"White" = "white",    # Replace these with the Flag categories and the colors you want to assign
"Green" = "green",
"Black" = "black",
"Yellow" = "yellow",  # Add more colors for other Flag categories as needed
"Red" = "red"
)) +
labs(
  title = "Marathon Performance by Flag",
  x = "Gender",
  y = "Marathon Performance (Runtimes)",
  fill = "Flag Color" # Title for the legend (Fill refers to the box color)
) +
theme(legend.position = "right")

# Print the plot
print(flag_exam)

```



```

flag_age_ranges<- ggplot(weather_parameters, aes(x = age_ranges, y = Runtimes, fill = Flag)) +
  geom_boxplot() + # Use geom_boxplot for creating a boxplot
  theme_bw() +     # Black and white theme for a clean plot
  scale_fill_manual(values = c(
    "White" = "white",    # Replace these with the Flag categories and the colors you want to assign
    "Green" = "green",
    "Black" = "black",
    "Yellow" = "yellow",  # Add more colors for other Flag categories as needed
    "Red" = "red"
  )) +

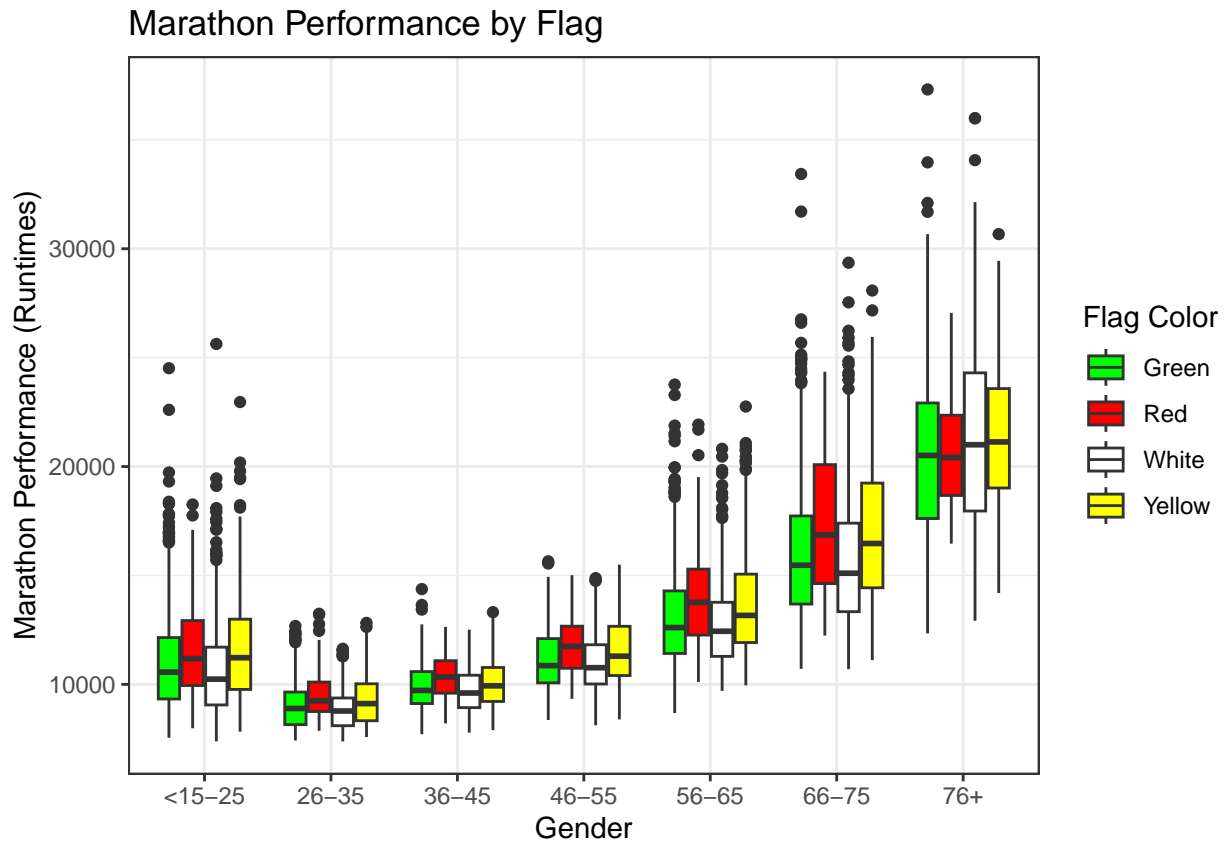
```

```

labs(
  title = "Marathon Performance by Flag",
  x = "Gender",
  y = "Marathon Performance (Runtimes)",
  fill = "Flag Color" # Title for the legend (Fill refers to the box color)
) +
theme(legend.position = "right")

# Print the plot
print(flag_age_ranges)

```



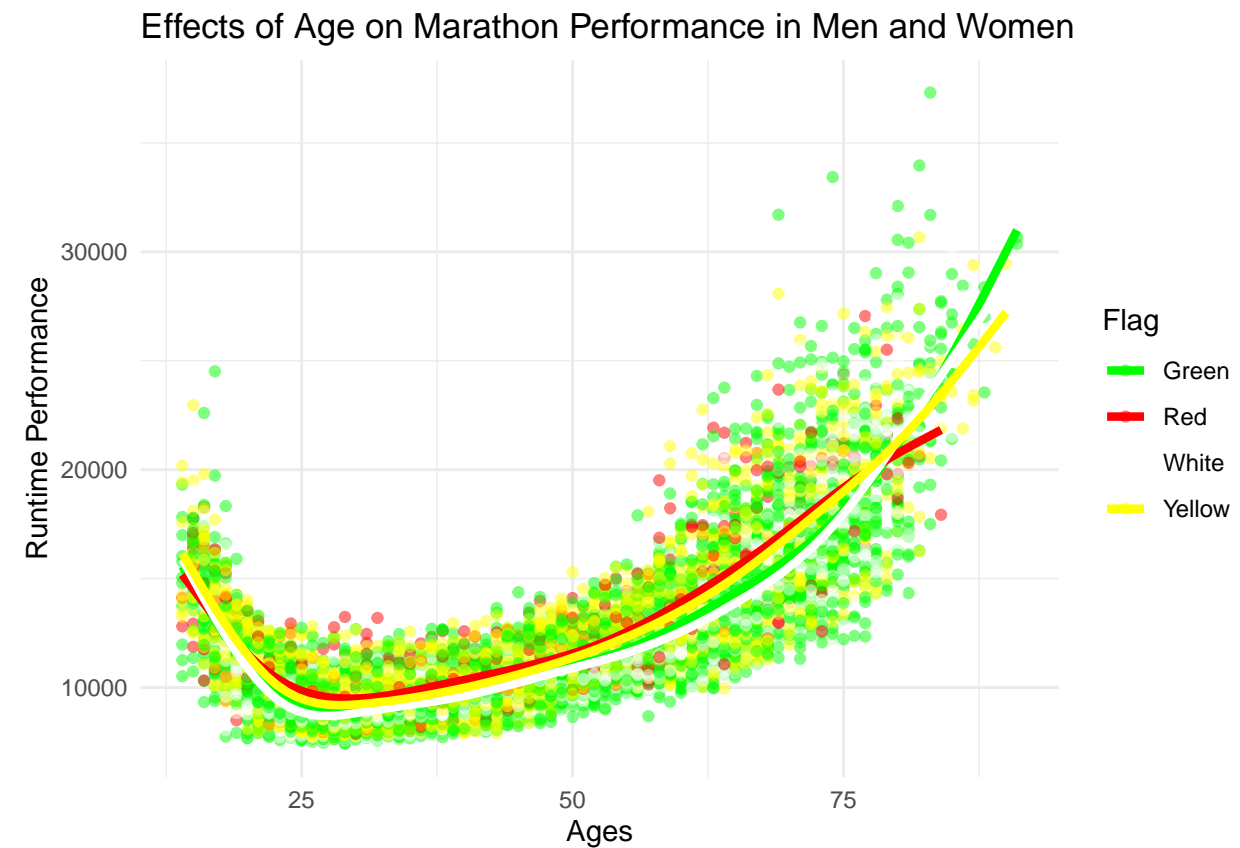
```

age_flag <- ggplot(course_record_project1, aes(x = `Age (yr)`, y = Runtimes, color = Flag)) +
  geom_point(alpha = 0.5) + # Set point transparency to 0.5 for better visibility
  geom_smooth(se = FALSE, linewidth = 1.5) + # Smooth line without confidence interval
  labs(
    title = "Effects of Age on Marathon Performance in Men and Women",
    x = "Ages",
    y = "Runtime Performance"
  ) +
  scale_color_manual(values = c(
    "White" = "white", # Custom color for each flag category
    "Green" = "green",
    "Black" = "black",
    "Yellow" = "yellow",
    "Red" = "red"
  )) +
  theme_minimal()

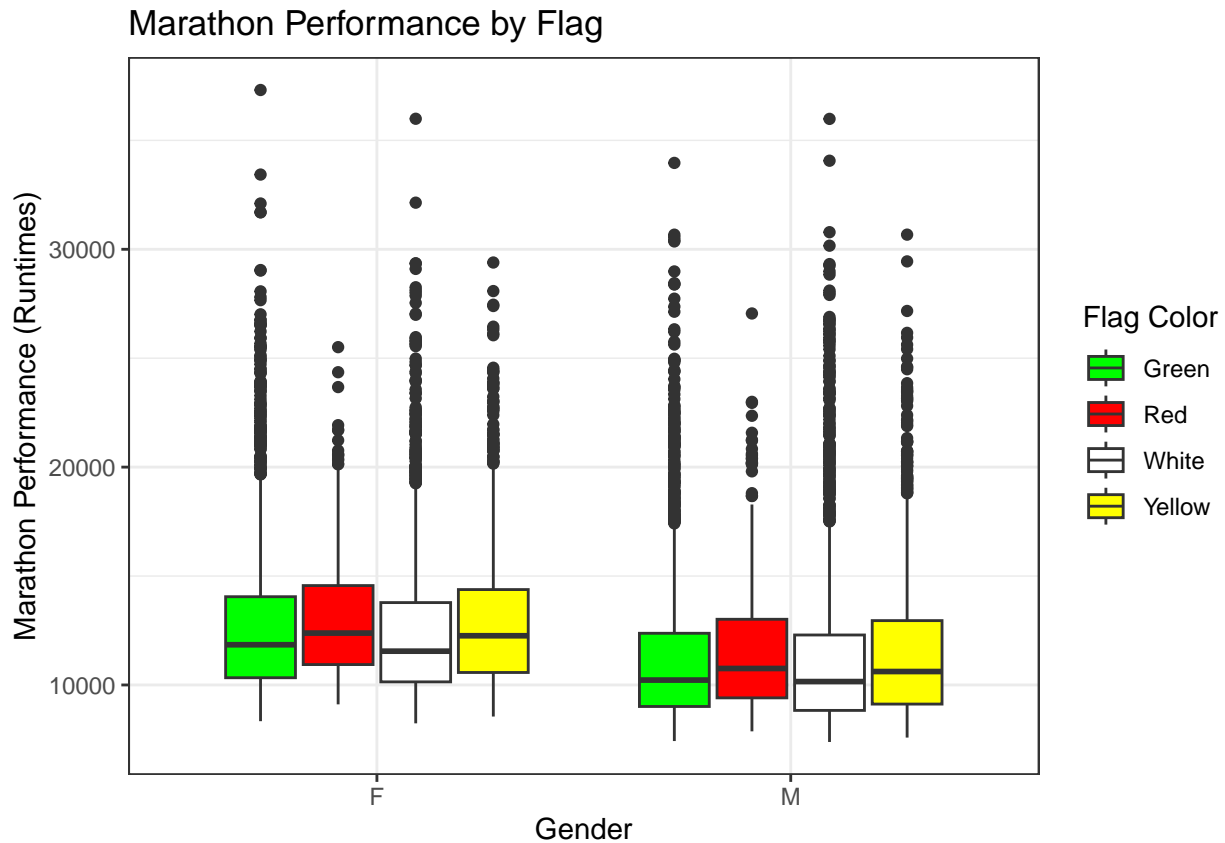
```

```
# Print the plot  
print(age_flag)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
print(flag_exam)
```

```
sum(course_record_project1$Flag=="Black")
```

```
## [1] 0
```

```
sum(course_record_project1$Flag=="Green")
```

```
## [1] 4706
```

```
sum(course_record_project1$Flag=="Red")
```

```
## [1] 592
```

```
sum(course_record_project1$Flag=="Yellow")
```

```
## [1] 2022
```

```
sum(course_record_project1$Flag=="White")
```

```
## [1] 3753
```

```
names(course_record_project1)
```

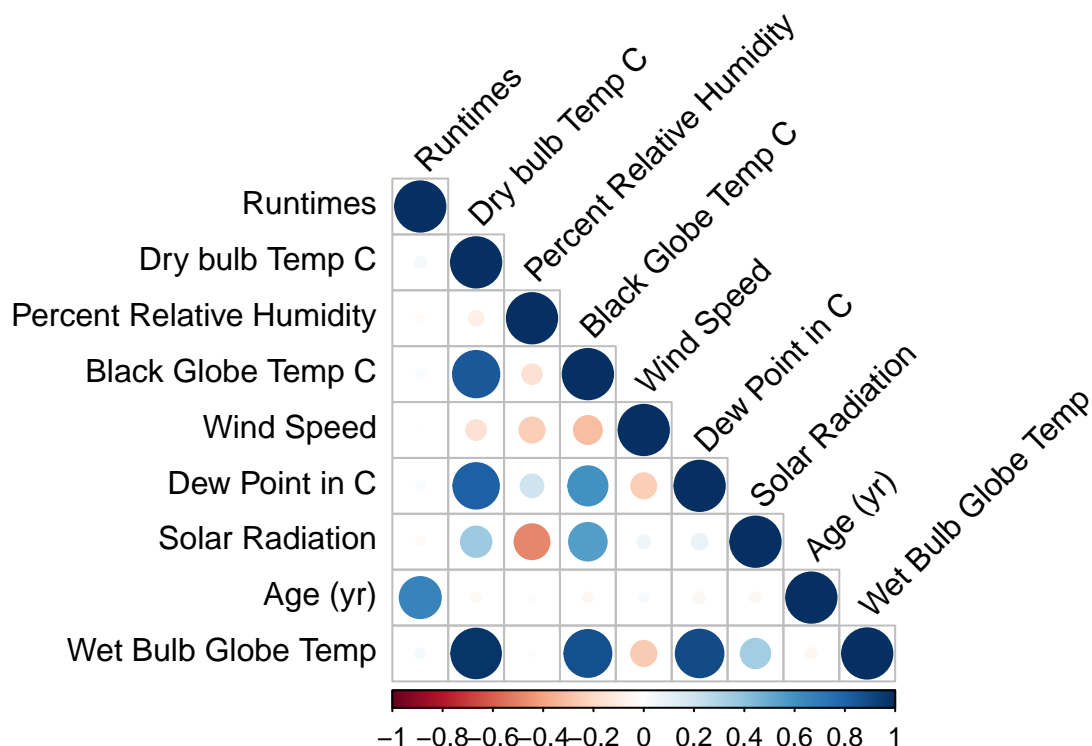
```
## [1] "Race" "Year"
## [3] "Gender" "Flag"
## [5] "Age (yr)" "Percent CR"
## [7] "Dry bulb Temp C" "Wet bulb Temp C"
## [9] "Percent Relative Humidity" "Black Globe Temp C"
## [11] "Solar Radiation" "Dew Point in C"
## [13] "Wind Speed" "Wet Bulb Globe Temp"
## [15] "CR" "Race_Seconds"
## [17] "Runtimes" "age_ranges"
```

```
## [19] "date"                                "avg_ppm"

# Step 1: Select only numeric columns from the data frame
airquality<- course_record_project1%>%
  select(Runtimes, `Dry bulb Temp C` , `Percent Relative Humidity` , `Black Globe Temp C` ,
    `Wind Speed` , `Dew Point in C` ,`Solar Radiation` , `Age (yr)` ,`Wet Bulb Globe Temp` , Year,
    avg_ppm)

# Step 2: Compute the correlation matrix
cor_matrix2 <- cor(airquality, use = "complete.obs") # Use complete.obs to ignore NAs

# Step 3: Visualize the correlation matrix using corrrplot
corrrplot(cor_matrix, method = "circle", type = "lower", tl.col = "black", tl.srt = 45)
```



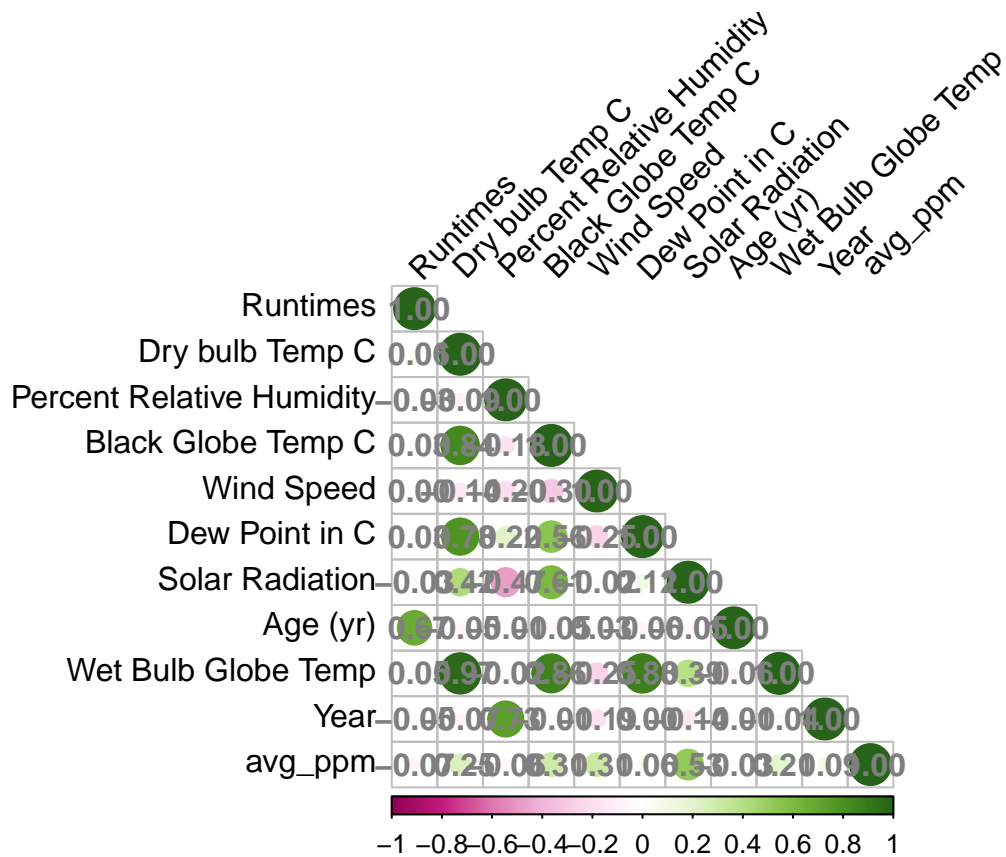
```
# Melt the correlation matrix for ggplot2
library(reshape2)
cor_data <- melt(cor_matrix2)

# Create the ggplot correlation heatmap
# Break down and Visualization of the correlation matrix with environmental on age variable names
corrrplot(cor_matrix2,
  method = "circle", # Use circle method for visualization
  type = "lower",    # Only display lower triangle
  tl.col = "black",  # Color of the text labels
  tl.srt = 45,       # Rotate the text labels at 45 degrees
  tl.pos = "lt",     # Display text labels on left and top
  col = COL2('PiYG'), # Color palette for the plot
  cl.pos = 'b',      # Place the color legend at the bottom
  addCoef.col = 'grey50', # Add correlation coefficient values in grey
```

```

is.corr = TRUE,          # Ensure this is treated as a correlation matrix
col.lim = c(-1, 1)      # Set the color limit for correlation values
)

```

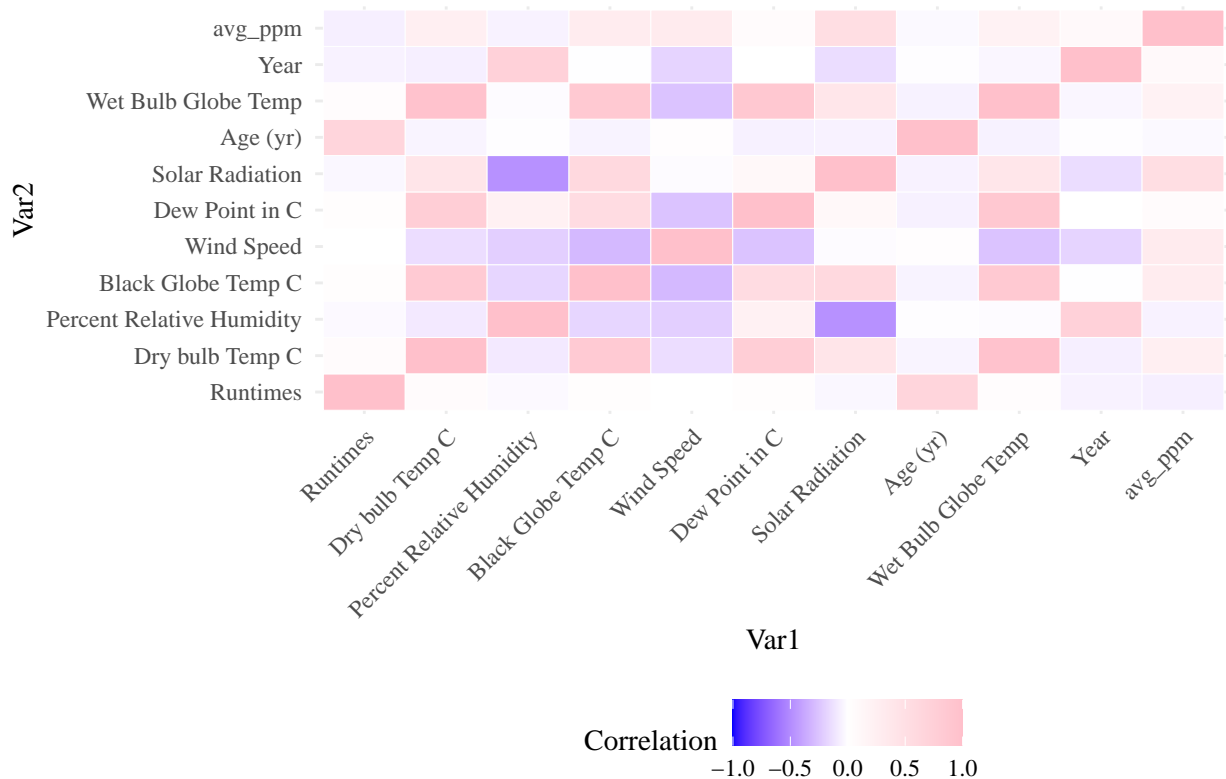


```

ggplot(data=cor_data, aes(x= Var1, y= Var2, fill=value))+
  geom_tile(color= "white")+
  scale_fill_gradient2(low= "blue", high="pink", mid = "white",
    midpoint=0,
    limit=c(-1,1),
    space="Lab",
    name="Correlation")+
  labs(title="Correlation Between Weather Variables and Performance")+
  theme_minimal(base_family="Times")+
  theme(axis.text.x= element_text(angle =45, vjust= 1, hjust = 1), legend.position="bottom")

```

Correlation Between Weather Variables and Performance



```
library(dplyr)

# Summarize data by Flag to calculate the median, IQR, Q1 (25th percentile), and Q3 (75th percentile) f
flag_summary <- course_record_project1 %>%
  group_by(Flag) %>%
  summarise(
    Median_Runtime = median(Runtimes, na.rm = TRUE),      # Median of Runtimes
    IQR_Runtime = IQR(Runtimes, na.rm = TRUE),           # IQR of Runtimes
    Q1_Runtime = quantile(Runtimes, 0.25, na.rm = TRUE),  # Lower quartile (25th percentile)
    Q3_Runtime = quantile(Runtimes, 0.75, na.rm = TRUE)   # Upper quartile (75th percentile)
  )

# Print the summary table
print(flag_summary)

## # A tibble: 4 x 5
##   Flag   Median_Runtime IQR_Runtime Q1_Runtime Q3_Runtime
##   <chr>         <dbl>         <dbl>     <dbl>     <dbl>
## 1 Green         10956.          3796.      9519.     13315.
## 2 Red           11659.          3809.     10100.     13909.
## 3 White         10817.          3720.      9386.     13106.
## 4 Yellow        11370.          3922.      9807.     13729.

# plot(model)
#
# cor(environmental_conditions)
# data_for_corr <- environmental_conditions %>%
#   select(`Dry bulb Temp C`, `Percent Relative Humidity`, `Black Globe Temp C`,
```

```

#       `Wind Speed`, `Dew Point in C`, `Solar Radiation`, `Age (yr)`,
#       `Wet Bulb Globe Temp`, Gender) %>%
#   mutate(Gender = as.numeric(as.factor(Gender)))
#
# library(corrplot)
# cor_matrix <- cor(data_for_corr, use = "complete.obs") # calculating correlation matrix
# corrplot(cor_matrix, method = "circle", type = "upper",
#          order = "hclust", tl.col = "black", tl.srt = 45)
#
#
#
#
#
# # Load necessary library
# library(dplyr)
#
# # Assuming environmental_conditions is your main dataset
# # Recreate the dataset making sure all variables are numeric
# data_for_corr <- environmental_conditions %>%
#   select(`Dry bulb Temp C`, `Percent Relative Humidity`, `Black Globe Temp C`,
#          `Wind Speed`, `Dew Point in C`, `Solar Radiation`, `Age (yr)`,
#          `Wet Bulb Globe Temp`, Gender) %>%
#   mutate(across(where(is.factor), as.numeric), # Convert all factors to numeric
#          across(where(is.character), as.numeric)) # Convert all characters to numeric
#
# # Check the structure of the data to confirm all are numeric
# str(data_for_corr)
#
# # Calculate the correlation matrix with complete observations
# # Assuming Gender is a factor with levels 'Male' and 'Female'
# data_for_corr$Gender <- as.numeric(data_for_corr$Gender) # Convert to 1 for Male, 0 for Female
# data_for_corr$Gender <- ifelse(data_for_corr$Gender == F, 0, 1)
# cor_matrix <- cor(data_for_corr, use = "complete.obs")
# str(data_for_corr)
#
# environmental_conditions <- course_record_project1 %>%
#   select(Race, Gender, `Age (yr)`, `Percent CRseconds`, `Dry bulb Temp C`, `Wet bulb Temp C`,
#          `Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`,
#          `Dew Point in C`, `Wind Speed`, `Wet Bulb Globe Temp`) %>%
#   group_by(Race, Gender, `Age (yr)`)
#
#
#
# age_boxplot <- ggplot(marathon_performance_by_age, aes(x = `Age (yr)` y = Percent_CRseconds, fill = Gender)) +
#   geom_boxplot() +
#   labs(title = "Effects of Age on Marathon Performance in Men and Women",
#        x = "Age Ranges",
#        y = "Percent CR Seconds (Performance)") +
#   scale_fill_manual(values = c("F" = "hotpink", "M" = "royalblue")) +
#   theme_minimal()
#
# ggplot(environmental_conditions, aes(x = x, y = y, color = group)) +
#   geom_point() + # Scatter plot

```

```
# stat_ellipse(level = 0.95) + # 95% confidence ellipse
# labs(title = "Scatter Plot with Ellipses",
#       x = "X Axis",
#       y = "Y Axis") +
# theme_minimal()
#
# library(ggplot2)
# library(dplyr)
#
names(course_record_project1)
```

```
## [1] "Race"                "Year"
## [3] "Gender"              "Flag"
## [5] "Age (yr)"            "Percent CR"
## [7] "Dry bulb Temp C"     "Wet bulb Temp C"
## [9] "Percent Relative Humidity" "Black Globe Temp C"
## [11] "Solar Radiation"     "Dew Point in C"
## [13] "Wind Speed"          "Wet Bulb Globe Temp"
## [15] "CR"                  "Race_Seconds"
## [17] "Runtimes"            "age_ranges"
## [19] "date"                "avg_ppm"
```

```
# # Filter and group the data as you described
environmental_conditions <- course_record_project1 %>%
  select(Race, Gender, `Age (yr)`, `Percent CR`, `Dry bulb Temp C`, `Wet bulb Temp C`,
         `Percent Relative Humidity`, `Black Globe Temp C`, `Solar Radiation`, `Dew Point in C`
#
# # Scatter plot with ellipses, using Dry bulb Temp and Wet Bulb Globe Temp as an example
# ggplot(environmental_conditions, aes(x = `Dry bulb Temp C`, y = `Wet Bulb Globe Temp`, color = Gender))
#   geom_point() + # Scatter plot
#   stat_ellipse(level = 0.95) + # 95% confidence ellipse
#   labs(title = "Scatter Plot of Environmental Conditions with Ellipses",
#         x = "Dry Bulb Temperature (°C)",
#         y = "Wet Bulb Globe Temperature (°C)") +
#   facet_wrap(~ Race) + # Facet by Race
#   theme_minimal()

# Scatter plot with ellipses, including Percent CR as point size
# ggplot(environmental_conditions, aes(x = `Dry bulb Temp C`, y = `Wet Bulb Globe Temp`, color = Gender))
#   geom_point(alpha = 0.7) + # Scatter plot with alpha for transparency
#   stat_ellipse(level = 0.95) + # 95% confidence ellipse
#   labs(title = "Scatter Plot of Environmental Conditions and Marathon Performance",
#         x = "Dry Bulb Temperature (°C)",
#         y = "Wet Bulb Globe Temperature (°C)",
#         size = "Percent CR (Performance)") + # Label for Percent CR
#   facet_wrap(~ Race) + # Facet by Race
#   theme_minimal()

# Explanation of Changes:
# aes(size = Percent CR): The Percent CR variable is mapped to the size of the points, so larger (or smaller) values
# alpha = 0.7: Adds transparency to the points, making overlapping points easier to visualize.
# labs(size = "Percent CR (Performance)")
```

```

# ggplot(envronmental_conditions, aes(x = `Dry bulb Temp C`, y =Percent_CRseconds, color = Gender)) +
#   geom_point() +
#   geom_labelsmooth(aes(label = Gender), fill = "white",
#                     method = "lm", formula = y ~ x,
#                     size = 3, linewidth = 1, boxlinewidth = 0.4) +
#   theme_bw() + guides(color = 'none') # remove legend
#
# library(ggplot2)
#
# # Scatter plot with linear smooth and gender labels
# ggplot(envronmental_conditions, aes(x = `Dry bulb Temp C`, y = Percent_CRseconds, color = Gender)) +
#   geom_point() + # Scatter plot
#   geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "black") + # Linear regression line
#   geom_label(aes(label = Gender), fill = "white", size = 3, color = "black") + # Adding labels with
#   theme_bw() + # Black and white theme
#   guides(color = 'none') + # Remove legend for color
#   labs(title = "Marathon Performance vs. Dry Bulb Temperature",
#        x = "Dry Bulb Temperature (°C)",
#        y = "Marathon Performance (Percent CR Seconds)")
#
# library(ggplot2)
#
# # Scatter plot with linear smooth, confidence interval, gender labels, and faceted by race
# ggplot(envronmental_conditions, aes(x = `Dry bulb Temp C`, y = Percent_CRseconds, color = Gender)) +
#   geom_point() + # Scatter plot
#   geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confi
#   # Adding labels with gender
#   facet_wrap(~ Race) + # Facet the plot by Race
#   theme_bw() + # Black and white theme
#   guides(color = 'none') + # Remove legend for color
#   labs(title = "Marathon Performance vs. Dry Bulb Temperature by Race",
#        x = "Dry Bulb Temperature (°C)",
#        y = "Marathon Performance (Percent CR Seconds)") +
#   scale_fill_manual(values = c("Male" = "blue", "Female" = "pink"))
#
# ggplot(envronmental_conditions, aes(x = `Dry bulb Temp C`, y = Percent_CRseconds, color = Gender)) +
#   geom_point() + # Scatter plot points
#   geom_smooth(method = "lm", formula = y ~ x, color = "black") + # Linear regression line with confi
#   facet_wrap(~ Race) + # Facet the plot by Race
#   theme_bw() + # Black and white theme
#   guides(color = 'none') + # Remove the legend for color
#   labs(title = "Marathon Performance vs. Dry Bulb Temperature by Race",
#        x = "Dry Bulb Temperature (°C)",
#        y = "Marathon Performance (Percent CR Seconds)") +
#   scale_fill_manual(values = c("Male" = "blue", "Female" = "pink")) # Custom colors for gender

```

```
sum(is.na(course_record_project1))
```

```
## [1] 4854
```

```

Missing_Data<- sapply(course_record_project1, function(x) sum(is.na(x)))
# Convert to dataframe
Missing_Data_df <- data.frame(ColumnNames = names(Missing_Data), `Missing Data` = Missing_Data)

# Set names for the dataframe columns if necessary
names(Missing_Data_df) <- c("Variables", "Missing Data")

as.data.frame(Missing_Data)

```

```

##                               Missing_Data
## Race                           0
## Year                           0
## Gender                         0
## Flag                           0
## Age (yr)                       0
## Percent CR                     0
## Dry bulb Temp C                0
## Wet bulb Temp C                0
## Percent Relative Humidity      0
## Black Globe Temp C             0
## Solar Radiation                0
## Dew Point in C                 0
## Wind Speed                     0
## Wet Bulb Globe Temp            0
## CR                             0
## Race_Seconds                  0
## Runtimes                       0
## age_ranges                    0
## date                          2427
## avg_ppm                       2427

```

```

Variables_table<- data_frame(
  Variables= c("Race", "Year", "Gender", "Flag", "Age (yr)",
    "Percent CR", "Dry bulb Temp C", "Wet bulb Temp C",
    "Percent Relative Humidity", "Black Globe Temp C", "Solar Radiation",
    "Dew Point in C", "Wind Speed", "Wet Bulb Globe Temp", "CR",
    "Race_Seconds", "Percent_CRseconds", "age_ranges"),
  Type= c("Character", "Numeric", "Character", "Character", "Numeric",
    "Numeric", "Numeric", "Numeric", "Numeric", "Numeric", "Numeric", "Numeric",
    "Numeric", "Numeric", "HMS/Numeric", "Numeric", "Numeric", "Categorical"),
  Description= c("Race represents the marathons the participants competed, including the B=Boston Marathon,
    C= Chicago Marathon, NY= New York City Marathon, T= Twin Cities Marathon (Minneapolis, MN),
    D= Grandma's Marathon (Duluth, MN).",
    "Years represented in the dataset ranging from 1993-2016.",
    "Gender is represented by F= Female and M= Male.",
    "Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C,
    Red= WBGT >23-28C, and Black= WBGT > 28C",
    "Age (yr) represents the ages of the participants.",
    "Percent CR is the percent off current course record for gender.",
    "Dry bulb Temp Celcius is the air temperature without taking into account of the humidity
    moisture.",
    "Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity.",
    "Percent Relative Humidity how much moisture is in the air compared to the maximum amount of
    moisture the air can hold.",
    "Black Globe Temp Celcius indicates how hot it feels in direct sunlight. It considers both the
    air temperature and the solar radiation."
)

```


Table 6: Marathon Runners' Data Description

Variables	Missing Data	Type	Description
Age (yr)	0	Numeric	Age (yr) represents the ages of the participants.
age_ranges	0	Categorical	Age_ranges are the age groups.
avg_ppm	2427	NA	NA
Black Globe Temp C	0	Numeric	Black Globe Temp Celcius indicates how hot the sun is.
CR	0	HMS/Numeric	CR is the course record for each marathon.
date	2427	NA	NA
Dew Point in C	0	Numeric	Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure).
Dry bulb Temp C	0	Numeric	Dry bulb Temp Celcius is the air temperature.
Flag	0	Character	Flag WBGT Thresholds. White= WBGT Thresholds.
Gender	0	Character	Gender is represented by F= Female and M= Male.
Percent CR	0	Numeric	Percent CR is the percent off current course record.
Percent Relative Humidity	0	Numeric	Percent Relative Humidity how much moisture is in the air.
Percent_CRseconds	NA	Numeric	Percent_CRseconds is the converted gender percentage into seconds.
Race	0	Character	Race represents the marathons the participants ran.
Race_Seconds	0	Numeric	Race_Seconds is the course record measured in seconds.
Runtimes	0	NA	NA
Solar Radiation	0	Numeric	Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels in the form of electromagnetic waves.
Wet Bulb Globe Temp	0	Numeric	Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed and solar radiation.
Wet bulb Temp C	0	Numeric	Wet bulb Temp Celcius is a measure of temperature and humidity.
Wind Speed	0	Numeric	Wind Speed in Km/hr.
Year	0	Numeric	Years represented in the dataset ranging from 1999 to 2014.

```

    "Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels in the form of electromagnetic waves.",
    "Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure).",
    "Wind Speed in Km/hr.",
    "Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed and solar radiation.",
    "CR is the course record for each marathon.",
    "Race_Seconds is the course record measured in seconds.",
    "Percent_CRseconds is the converted gender percentage into seconds.",
    "Age_ranges are the age groups."
  )
Missing_Data_df$Variables <- as.character(Missing_Data_df$Variables)
Variables_table$Variables <- as.character(Variables_table$Variables)
merged_df <- merge(Missing_Data_df, Variables_table, by = "Variables", all = TRUE)

merged_df%>%
  kbl(caption = "Marathon Runners' Data Description") %>%
  kable_classic(full_width = F, html_font = "Cambria", font_size= 12)

# Create the table with kable and customize with kableExtra
kable(merged_df, "latex", booktabs = TRUE, caption = "Marathon Runners' Data Description") %>%
  kable_styling(latex_options = c("striped", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%

```

```

column_spec(4, width = "8cm")

library(knitr)
library(kableExtra)

# Assuming 'merged_df' is already loaded and contains your data
# Create the table using kable and customize with kableExtra

kable(merged_df, "latex", booktabs = TRUE, longtable = TRUE, caption = "Marathon Runners' Data Descripti
  kable_styling(latex_options = c("striped", "scale_down", "hold_position"), full_width = FALSE) %>%
  column_spec(1, width = "3cm") %>% # Adjust width based on content
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "3cm") %>%
  column_spec(4, width = "10cm")

## Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
## resized.

```

Table 8: Marathon Runners' Data Description

Variables	Missing Data	Type	Description
Age (yr)	0	Numeric	Age (yr) represents the ages of the participants.
age_ranges	0	Categorical	Age_ranges are the age groups.
avg_ppm	2427	NA	NA
Black Globe Temp C	0	Numeric	Black Globe Temp Celcius indicates how hot it feels in direct sunlight.It considers temperature, humidity, wind speed, sun angle, and cloud cover to provide a holistic view of the stress placed on the body in hot environments.
CR	0	HMS/Numeric	CR is the course record for each marathon.
date	2427	NA	NA
Dew Point in C	0	Numeric	Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form. If the air were to be cooled even more, water vapor would have to come out of the atmosphere in the liquid form, usually as fog or precipitation.
Dry bulb Temp C	0	Numeric	Dry bulb Temp Celcius is the air temperature without taking in account of the humidity or any moisture.
Flag	0	Character	Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C, Red= WBGT >23-28C, and Black= WBGT > 28C
Gender	0	Character	Gender is represented by F= Female and M= Male.
Percent CR	0	Numeric	Percent CR is the percent off current course record for gender.
Percent Relative Humidity	0	Numeric	Percent Relative Humidity how much moisture is in the air compared to the maximum amount of moisture the air can hold at a given temperature. Gives an idea of how humid it feels outside.
Percent_CRseconds	NA	Numeric	Percent_CRseconds is the converted gender percentage into seconds.
Race	0	Character	Race represents the marathons the participants competed, including the B=Boston Marathon, C= Chicago Marathon, NY= New York City Marathon,T= Twin Cities Marathon (Minneapolis,MN), D= Grandma's Marathon (Duluth, MN).
Race_Seconds	0	Numeric	Race_Seconds is the course record measured in seconds.

Runtimes	0	NA	NA
Solar Radiation	0	Numeric	Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels through space and reaches the Earth as light and heat.
Wet Bulb Globe Temp	0	Numeric	Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar radiation on humans. Formula WBGT= 0.7 x Tw + 0.2 x Tg+ 0.1 xTd.
Wet bulb Temp C	0	Numeric	Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity in the air. Wet bulb temperature gives you an idea of how temperature feels when you take humidity into account.
Wind Speed	0	Numeric	Wind Speed in Km/hr.
Year	0	Numeric	Years represented in the dataset ranging from 1993-2016.

#Aim 3 what has the largest impact

```
library(ggwordcloud)
library(ggplot2)
library(magick)

## Linking to ImageMagick 6.9.12.93
## Enabled features: cairo, fontconfig, freetype, heic, lcms, pango, raw, rsvg, webp
## Disabled features: fftw, ghostscript, x11

# Path to your image
fig_path <- "/Users/diahminhawkins/Documents/GitHub/Project1/weather.png"
getwd()

## [1] "/Users/diahminhawkins/Documents/GitHub/Project1"

# Load the image using magick
img <- image_read(fig_path)

# Convert image to raster for use in ggplot
img_raster <- as.raster(img)

# Example data
words <- c("Boston", "New York City", "Minneapolis", "Grandma's", "Chicago",
           "Race", "Age", "Gender", "Weather", "Performance",
           "Wet Bulb Globe Temperature", "Humidity")

frequencies <- c(1, 1, 1, 1, 1, 1, 5, 1, 4, 15, 3, 14)

new_frame <- data.frame(words, frequencies)

# Generate the word cloud on top of the image background
ggplot(new_frame, aes(label = words, size = frequencies)) +
  # Add the image background
  annotation_raster(img_raster, xmin = -Inf, xmax = Inf, ymin = -Inf, ymax = Inf) +

  # Generate the word cloud
  geom_text_wordcloud(aes(color = frequencies)) +
  scale_size_area(max_size = 20) +
```

Table 7: Marathon Runners' Data Description

Variables	Missing Data	Type	Description
Age (yr)	0	Numeric	Age (yr) represents the ages of the participants.
age_ranges	0	Categorical	Age_ranges are the age groups.
avg_ppm	2427	NA	NA
Black Globe Temp C	0	Numeric	Black Globe Temp Celcius indicates how hot it feels in direct sunlight.It considers temperature, humidity, wind speed, sun angle, and cloud cover to provide a holistic view of the stress placed on the body in hot environments.
CR	0	HMS/Numeric	CR is the course record for each marathon.
date	2427	NA	NA
Dew Point in C	0	Numeric	Dew Point in Celcius is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form. If the air were to be cooled even more, water vapor would have to come out of the atmosphere in the liquid form, usually as fog or precipitation.
Dry bulb Temp C	0	Numeric	Dry bulb Temp Celcius is the air temperature without taking into account of the humidity or any moisture.
Flag	0	Character	Flag WBGT Thresholds. White= WBGT < 10C, Green= WBGT 10-18C, Yellow=WBGT >18-23C, Red= WBGT >23-28C, and Black= WBGT > 28C
Gender	0	Character	Gender is represented by F= Female and M= Male.
Percent CR	0	Numeric	Percent CR is the percent off current course record for gender.
Percent Relative Humidity	0	Numeric	Percent Relative Humidity how much moisture is in the air compared to the maximum amount of moisture the air can hold at a given temperature. Gives an idea of how humid it feels outside.
Percent_CRseconds	NA	Numeric	Percent_CRseconds is the converted gender percentage into seconds.
Race	0	Character	Race represents the marathons the participants competed, including the B=Boston Marathon, C= Chicago Marathon, NY= New York City Marathon,T= Twin Cities Marathon (Minneapolis,MN), D= Grandma's Marathon (Duluth, MN).
Race_Seconds	0	Numeric	Race_Seconds is the course record measured in seconds.
Runtimes	0	NA	NA
Solar Radiation	0	Numeric	Solar Radiation in Watts per meter squared is the energy emitted by the sun, which travels through space and reaches the Earth as light and heat.
Wet Bulb Globe Temp	0	Numeric	Wet Bulb Globe Temp measures the combined effect of temperature, humidity, wind speed, and solar radiation on humans. Formula WBGT= $0.7 \times Tw + 0.2 \times Tg + 0.1 \times Td$.
Wet bulb Temp C	0	Numeric	Wet bulb Temp Celcius is a measure of temperature that reflects both the heat and humidity in the air. Wet bulb temperature gives you an idea of how temperature feels when you take humidity into account.
Wind Speed	0	Numeric	Wind Speed in Km/hr.
Year	0	Numeric	Years represented in the dataset ranging from 1992-2016

```
# Customize the colors of the words
scale_color_gradient(low = "yellow", high = "red") +

# Remove axis titles and labels since we want the word cloud only
theme_void()
```

