Hw4_b-d

2024-09-26

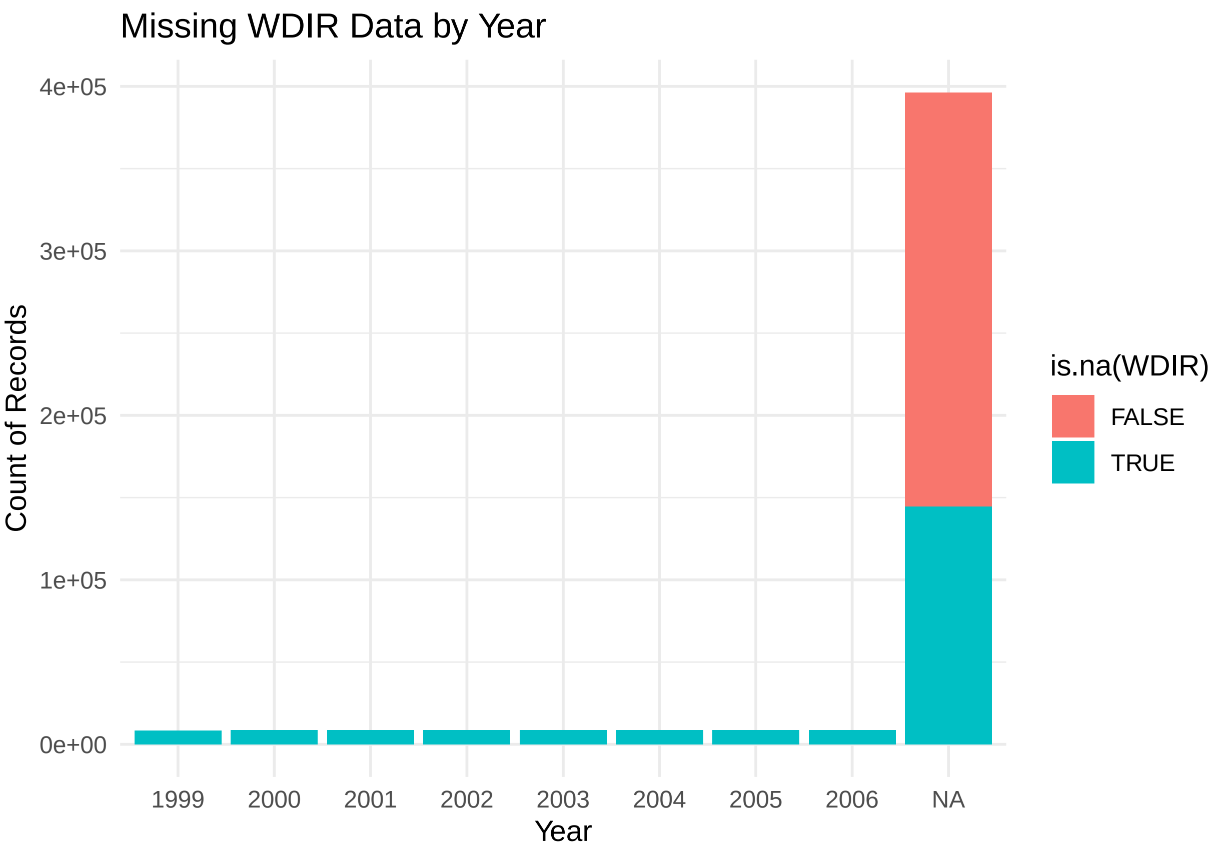
```
if (!exists("all_buoy_data")) {
  all_buoy_data <- fread("all_buoy_data.csv")
}
```

```
library(data.table)
library(ggplot2)
library(dplyr)
```

```
# Function to read buoy data for a specific year
read_buoy_data <- function(year) {
  file_root <- "https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
  tail <- ".txt.gz&dir=data/historical/stdmet/"
  path <- paste0(file_root, year, tail)

  # Printing the current year being processed
  print(paste("Processing year:", year))

  # Use fread to read in the data
  header <- scan(path, what = 'character', nlines = 1)
  buoy <- fread(path, header = FALSE, skip = 1, fill = TRUE)

  colnames(buoy) <- header
  buoy$date <- make_date(buoy$YYYY, buoy$MM, buoy$DD)
  return(buoy)
}
```

B:

```
# Converting 999 values to NA for the relevant columns
convert_to_na <- function(data) {
  data %>%
    mutate(across(c(WDIR, WVHT, MWD, BAR, ATMP, WTMP, DEWP, VIS), ~ na_if(., 999)))
}

# Applying the conversion to the buoy data
clean_buoy_data <- convert_to_na(all_buoy_data)

# Summarizing NA patterns
na_summary <- clean_buoy_data %>%
  summarise(across(everything(), ~ sum(is.na(.)), .names = "na_{col}"))

# Visualizing the pattern of missing data over time (using year as a factor)
ggplot(clean_buoy_data, aes(x = as.factor(YYYY), fill = is.na(WDIR))) +
  geom_bar() +
  labs(title = "Missing WDIR Data by Year", x = "Year", y = "Count of Records") +
  theme_minimal()
```



Explanation for B: In the above code, we converted the values that are 999 to NA. Then we analyzed the missing data using a summary of NA counts across the dataset. The columns where 999 was replaced showed varying amounts of missing data.

Most of the years (1985 to 2006) have a very small amount of missing WDIR data. There's an increase in missing data for WDIR in more recent years. The bar labeled "NA" represents records without a year (perhaps incomplete or incorrect entries), which contain a significant amount of missing data. We can probably say that there is more missing data in recent years because of improper data collection.

Is it always appropriate to convert missing/null data to NA's? If the data is like the one we used and 999 is being used as a placeholder instead of NA, then converting them to NA makes sense as it would be handled correctly.

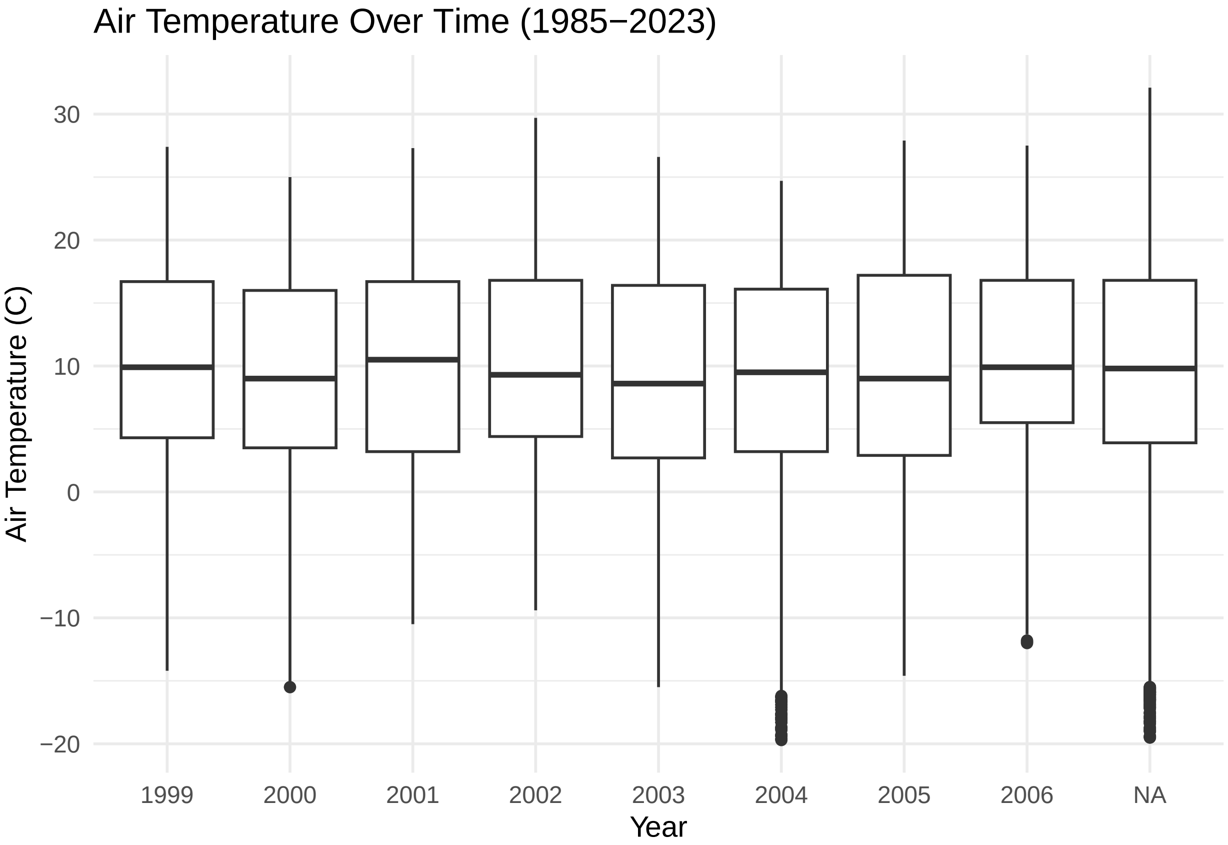
When might it not be? It might not be appropriate if the data means something. If 999 actually implied something like if 999 meant the equipment isn't functioning then it might not be appropriate to change them to NA.

Analyze the pattern of NA's. Do you spot any patterns in the way/dates that these are distributed? From the bar graph, we can see that the NAs are concentrated mainly in the recent years.

C:

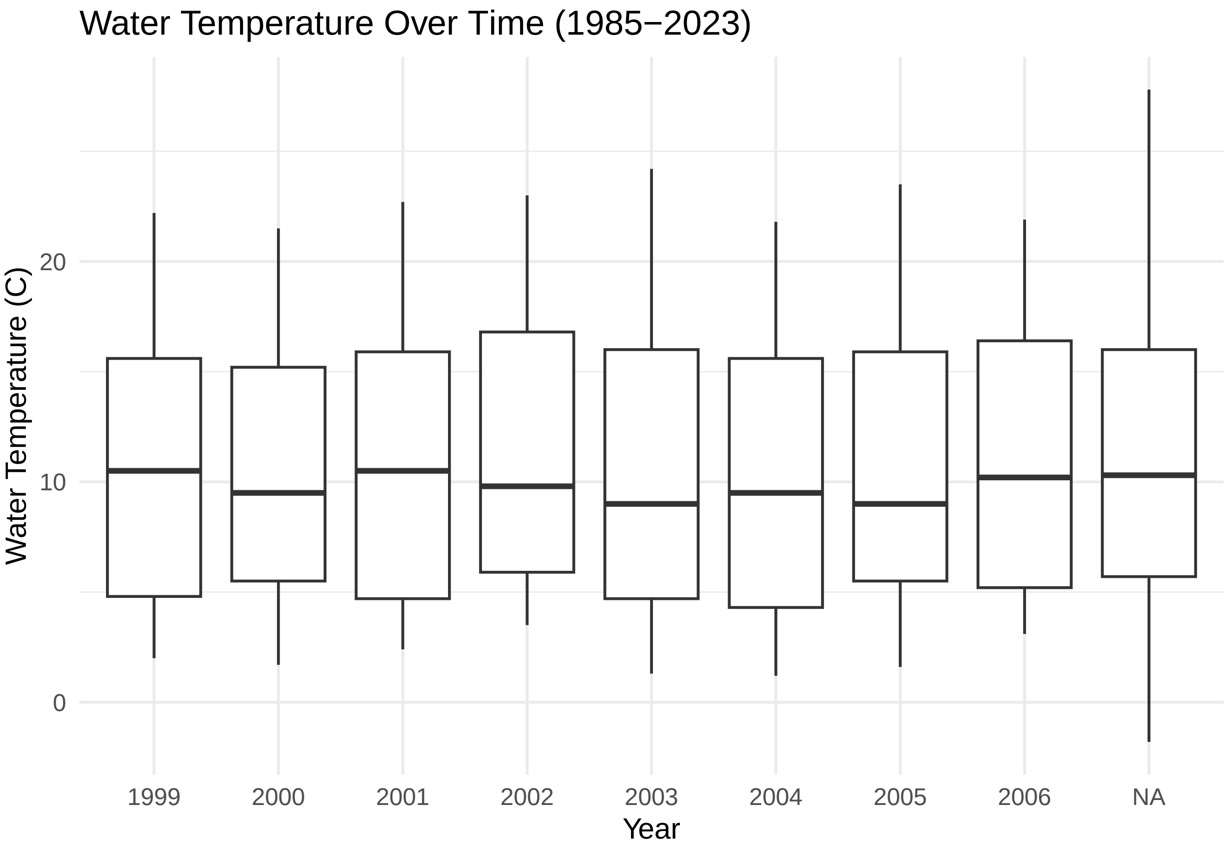
```
library(ggplot2)
library(dplyr)
```

```r
# Filtering out rows with NA values for key climate-related variables
climate_data <- clean_buoy_data %>%
  filter(!is.na(ATMP) & !is.na(WTMP) & !is.na(WSPD))

# Visualizing the trend in Air Temperature over time
ggplot(climate_data, aes(x = as.factor(YYYY), y = ATMP, group = YYYY)) +
  geom_boxplot() +
  labs(title = "Air Temperature Over Time (1985-2023)", x = "Year", y = "Air Temperature (C)") +
  theme_minimal()
```



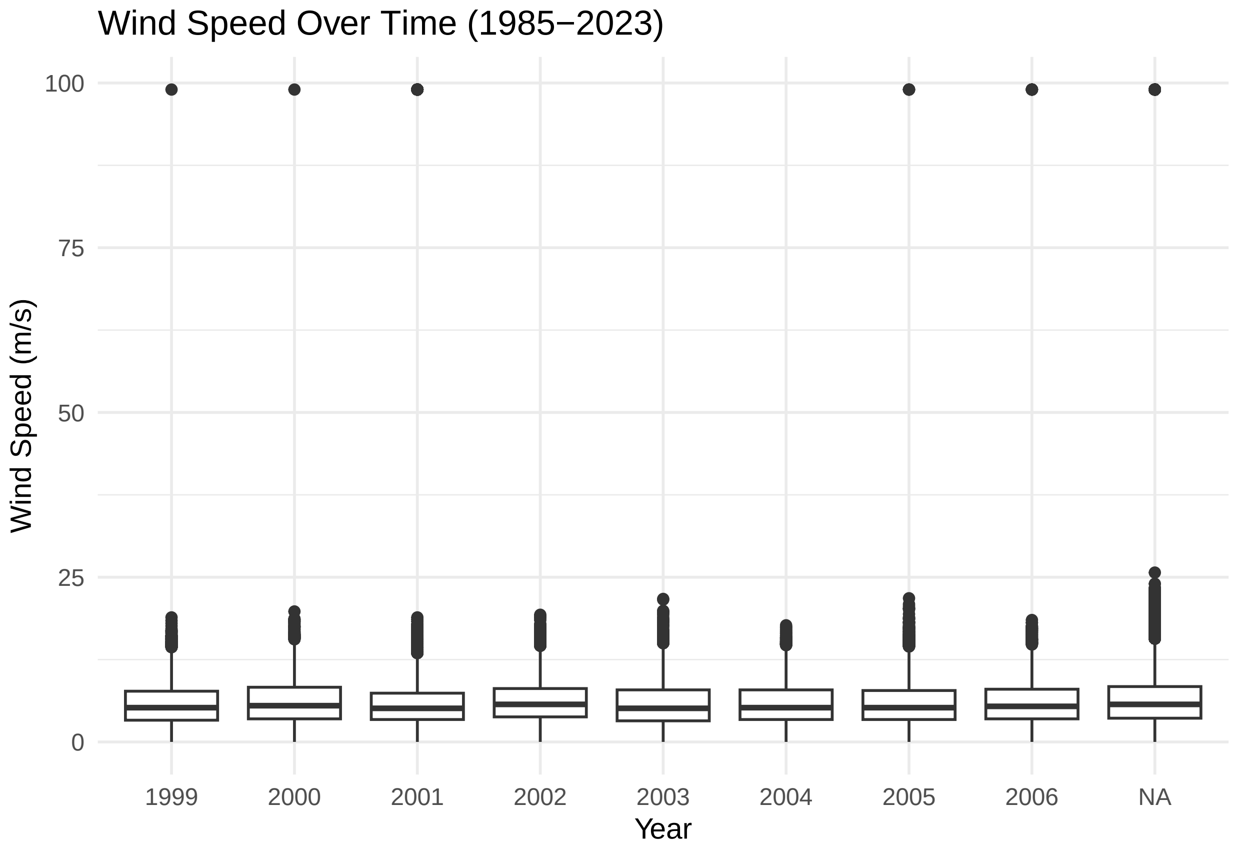
There is no significant long-term trend is observed in air temperature. There are minor variations over the years, but the temperature is pretty stable over time.

```r
ggplot(climate_data, aes(x = as.factor(YYYY), y = WTMP, group = YYYY)) +
  geom_boxplot() +
  labs(title = "Water Temperature Over Time (1985-2023)", x = "Year", y = "Water Temperature (C)") +
  theme_minimal()
```



Similar to air temperature, there doesn't seem to be any significant long-term trend in water temperature over the years.

```r
ggplot(climate_data, aes(x = as.factor(YYYY), y = WSPD, group = YYYY)) +
  geom_boxplot() +
  labs(title = "Wind Speed Over Time (1985-2023)", x = "Year", y = "Wind Speed (m/s)") +
  theme_minimal()
```



Wind speed exhibits more variability compared to air and water temperature, with some extreme values. But, the median wind speeds remain consistent over time, and no clear upward or downward trend is visible.

```r
# Optionally, calculating statistical summaries to support the visualization
climate_summary <- climate_data %>%
  group_by(YYYY) %>%
  summarise(
    avg_air_temp = mean(ATMP, na.rm = TRUE),
    avg_water_temp = mean(WTMP, na.rm = TRUE),
    avg_wind_speed = mean(WSPD, na.rm = TRUE)
  )

print(climate_summary)
```

```
## # A tibble: 9 x 4
##    YYYY avg_air_temp avg_water_temp avg_wind_speed
##   <int>        <dbl>          <dbl>          <dbl>
## 1  1999        10.0           10.3            5.73
## 2  2000         8.83          10.0            6.12
## 3  2001        10.0           10.6            6.60
## 4  2002        10.2           11.0            6.17
## 5  2003         8.80          10.0            5.83
## 6  2004         8.91           9.97           5.83
## 7  2005         9.32          10.2            5.87
## 8  2006        10.3           10.8            6.04
## 9    NA         9.92          10.9            9.31
```

Final Conclusion for C: From the visualizations and summaries, we do not observe a clear upward or downward trend in air temperature, water temperature, or wind speed over the time. All of these variables show pretty consistent behavior, with a few minor fluctuations in their averages and ranges across the years. Wind speed is the most variable of the three, but even here, the overall trend is stable.

D:

```
#Reading the Rainfall data
rainfall_data <- read.csv("Rainfall.csv")

str(rainfall_data)
```

```
## ’data.frame':    31714 obs. of  6 variables:
##  $ STATION         : chr  "COOP:190770" "COOP:190770" "COOP:190770" "COOP:190770" ...
##  $ STATION_NAME    : chr  "BOSTON LOGAN INTERNATIONAL AIRPORT MA US" "BOSTON LOGAN INTERNATIONAL AIR
##  $ DATE            : chr  "19850101 01:00" "19850101 09:00" "19850101 10:00" "19850101 11:00" ...
##  $ HPCP            : num  0 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 ...
##  $ Measurement.Flag: chr  "g" " " " " " " " " " " ...
##  $ Quality.Flag    : logi  NA NA NA NA NA NA ...
```

```
summary(rainfall_data)
```

```
##    STATION          STATION_NAME          DATE                HPCP        Measurement.Flag   Quali
##  Length:31714       Length:31714       Length:31714       Min.   :0.00000   Length:31714       Mode:
##  Class :character   Class :character   Class :character   1st Qu.:0.00000   Class :character   NA's:
##  Mode  :character   Mode  :character   Mode  :character   Median :0.01000   Mode  :character
##                                                           Mean   :0.03875
##                                                           3rd Qu.:0.04000
##                                                           Max.   :2.03000
```

Key Columns: STATION: The station code for where the data was collected. STATION_NAME: The full name of the station (Boston Logan International Airport). DATE: The date and time of the rainfall measurement. HPCP: Represents rainfall? Measurement.Flag and Quality.Flag: These flags might indicate measurement issues or quality concerns.

Cleaning the data:

```
# Converting DATE to a proper date-time format
rainfall_data$DATE <- as.POSIXct(rainfall_data$DATE, format = "%Y%m%d %H:%M")

# Filtering the data to include years only from 1985 to 2013
rainfall_data <- rainfall_data %>%
  filter(format(DATE, "%Y") >= 1985 & format(DATE, "%Y") <= 2013)

# Handling missing rainfall values (HPCP)
rainfall_data <- rainfall_data %>%
  filter(!is.na(HPCP))

# Adding an year column for merging
rainfall_data$YYYY <- as.integer(format(rainfall_data$DATE, "%Y"))
```

```r
# Checking unique years in both datasets
unique_years_buoy <- length(unique(clean_buoy_data$YYYY))
unique_years_rainfall <- length(unique(rainfall_data$YYYY))

# Checking if there are duplicate years
duplicate_buoy_years <- clean_buoy_data %>%
  group_by(YYYY) %>%
  summarise(count = n()) %>%
  filter(count > 1)

duplicate_rainfall_years <- rainfall_data %>%
  group_by(YYYY) %>%
  summarise(count = n()) %>%
  filter(count > 1)

# Printing out the results
print(duplicate_buoy_years)
```

```
## # A tibble: 9 x 2
##    YYYY  count
##   <int>  <int>
## 1  1999   8348
## 2  2000   8784
## 3  2001   8760
## 4  2002   8760
## 5  2003   8759
## 6  2004   8761
## 7  2005   8708
## 8  2006   8723
## 9    NA 396370
```

```r
print(duplicate_rainfall_years)
```

```
## # A tibble: 29 x 2
##     YYYY count
##    <int> <int>
##  1  1985   707
##  2  1986   692
##  3  1987   759
##  4  1988   600
##  5  1989   793
##  6  1990   723
##  7  1991   691
##  8  1992   743
##  9  1993   764
## 10  1994   817
## # i 19 more rows
```

```r
# Summarizing rainfall data by year- total rainfall per year
rainfall_yearly <- rainfall_data %>%
  group_by(YYYY) %>%
  summarise(total_rainfall = sum(HPCP, na.rm = TRUE), avg_rainfall = mean(HPCP, na.rm = TRUE))
```

```r
# Checking the summarized rainfall data
head(rainfall_yearly)
```

```
## # A tibble: 6 x 3
##    YYYY total_rainfall avg_rainfall
##   <int>          <dbl>        <dbl>
## 1  1985           36.6       0.0517
## 2  1986           44.3       0.0640
## 3  1987           45.2       0.0596
## 4  1988           34.8       0.0580
## 5  1989           42.4       0.0535
## 6  1990           46.5       0.0643
```

```r
# Summarizing buoy data by year
buoy_yearly <- clean_buoy_data %>%
  group_by(YYYY) %>%
  summarise(
    avg_air_temp = mean(ATMP, na.rm = TRUE),
    avg_water_temp = mean(WTMP, na.rm = TRUE),
    avg_wind_speed = mean(WSPD, na.rm = TRUE)
  )

# Checking the summarized buoy data
head(buoy_yearly)
```

```
## # A tibble: 6 x 4
##    YYYY avg_air_temp avg_water_temp avg_wind_speed
##   <int>        <dbl>          <dbl>          <dbl>
## 1  1999        10.0           10.3            6.70
## 2  2000         8.83          10.0            6.51
## 3  2001        10.0           10.6            7.20
## 4  2002        10.2           11.0            6.51
## 5  2003         8.79          10.0            6.17
## 6  2004         8.91           9.97           5.83
```

Merging the data:

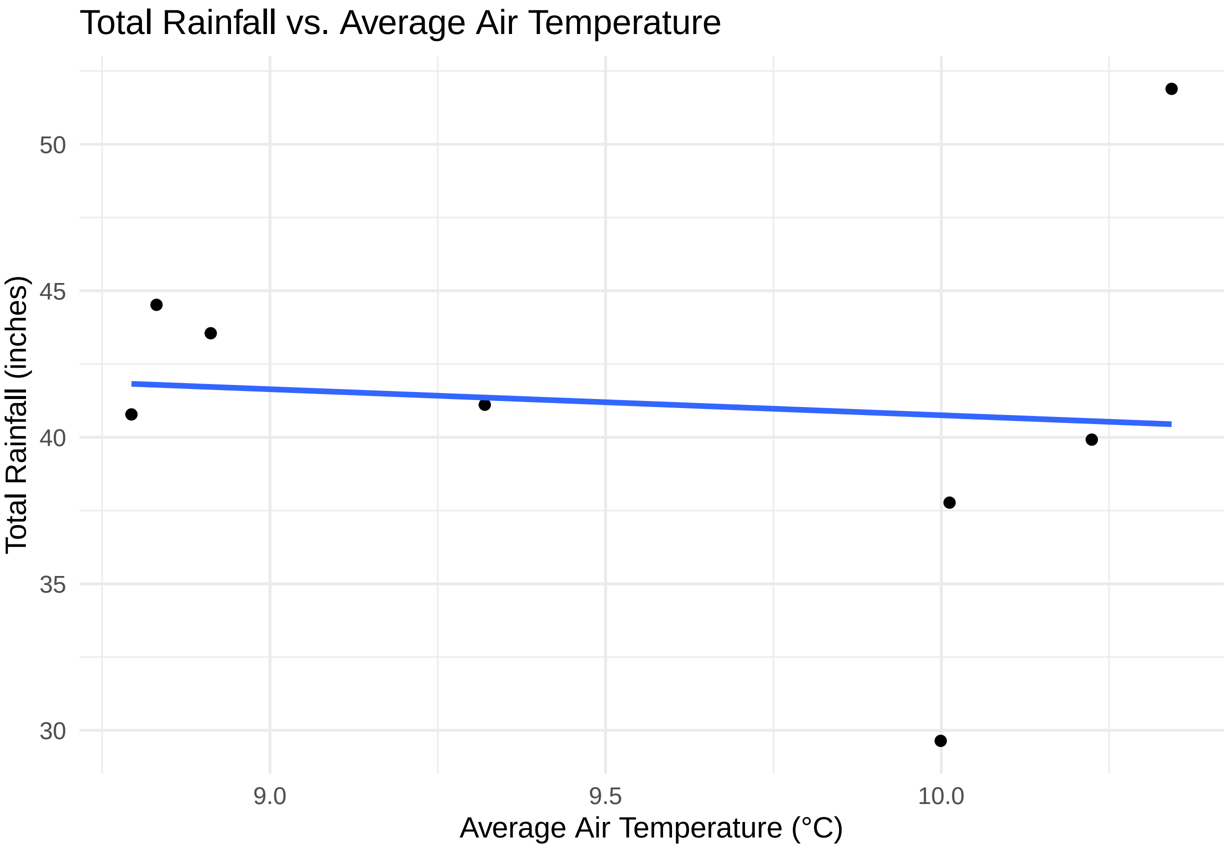
```r
# Merging summarized buoy and rainfall data by year
merged_data <- merge(buoy_yearly, rainfall_yearly, by = "YYYY")

# Checking the merged data
head(merged_data)
```

```
##   YYYY avg_air_temp avg_water_temp avg_wind_speed total_rainfall avg_rainfall
## 1 1999    10.012326      10.283880       6.703630          37.77   0.03009562
## 2 2000     8.831031      10.007136       6.508117          44.52   0.03111111
## 3 2001     9.999126      10.613174       7.197237          29.64   0.02752089
## 4 2002    10.224172      11.046703       6.511370          39.92   0.02909621
## 5 2003     8.793718       9.999621       6.166526          40.78   0.02764746
## 6 2004     8.911838       9.967059       5.834859          43.55   0.03391745
```

Rainfall vs Temperature:

```r
ggplot(merged_data, aes(x = avg_air_temp, y = total_rainfall)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Total Rainfall vs. Average Air Temperature", x = "Average Air Temperature (°C)", y = "Te
  theme_minimal()
```

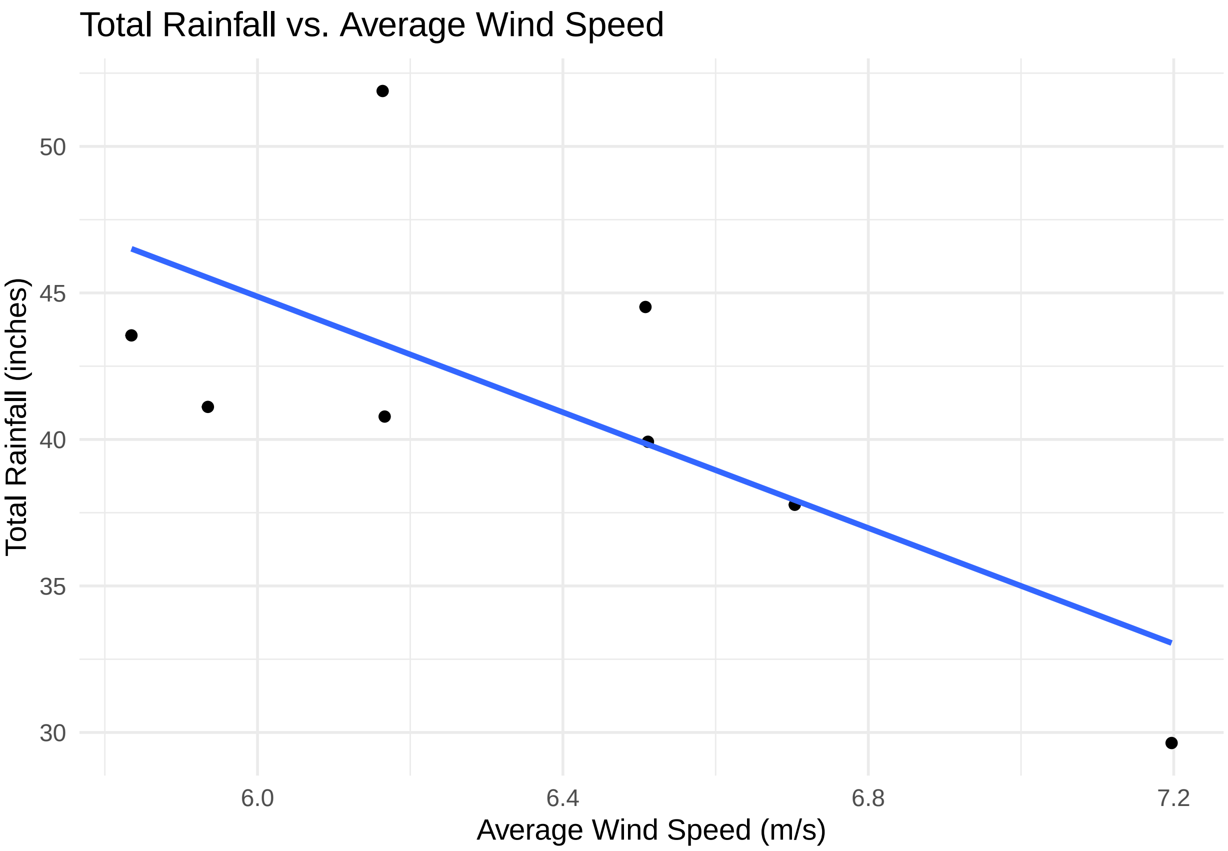
```
## 'geom_smooth()' using formula = 'y ~ x'
```



Rainfall vs wind speed:

```r
ggplot(merged_data, aes(x = avg_wind_speed, y = total_rainfall)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Total Rainfall vs. Average Wind Speed", x = "Average Wind Speed (m/s)", y = "Total Rain
  theme_minimal()
```

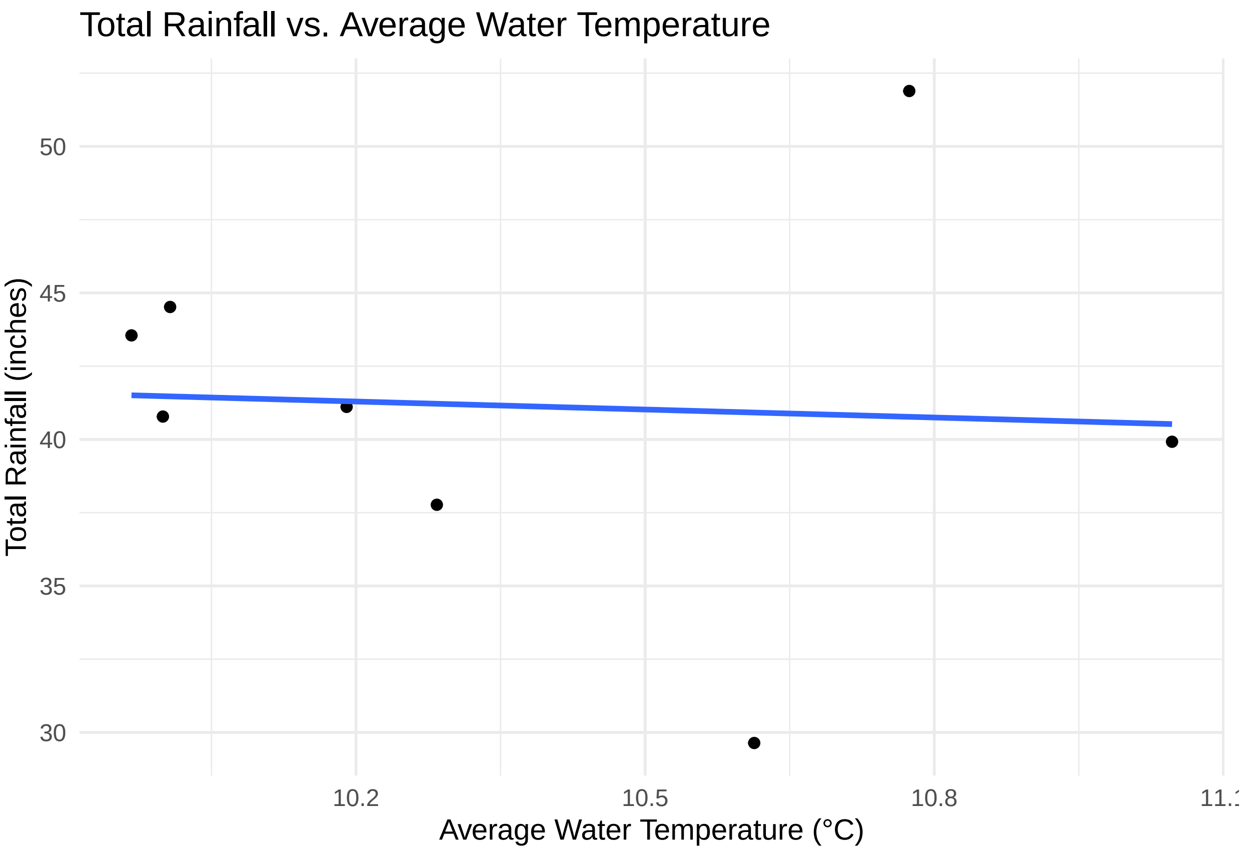
```
## 'geom_smooth()' using formula = 'y ~ x'
```



Rainfall vs Water temperature:

```
ggplot(merged_data, aes(x = avg_water_temp, y = total_rainfall)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Total Rainfall vs. Average Water Temperature", x = "Average Water Temperature (°C)", y =
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observations: Total Rainfall vs. Average Air Temperature: The plot shows a slight downward trend in rainfall as air temperature increases. This suggests that higher air temperatures could correspond with slightly less rainfall, but the correlation looks weak.

Total Rainfall vs. Average Wind Speed: The plot for rainfall and wind speed shows a more pronounced negative correlation, where higher wind speeds tend to correlate with lower rainfall. This could indicate that areas with stronger winds experience less rainfall.

Total Rainfall vs. Average Water Temperature: The relationship between water temperature and rainfall is nearly flat, suggesting almost no correlation between these two variables in this dataset. It seems that water temperature does not really influence the total rainfall here.

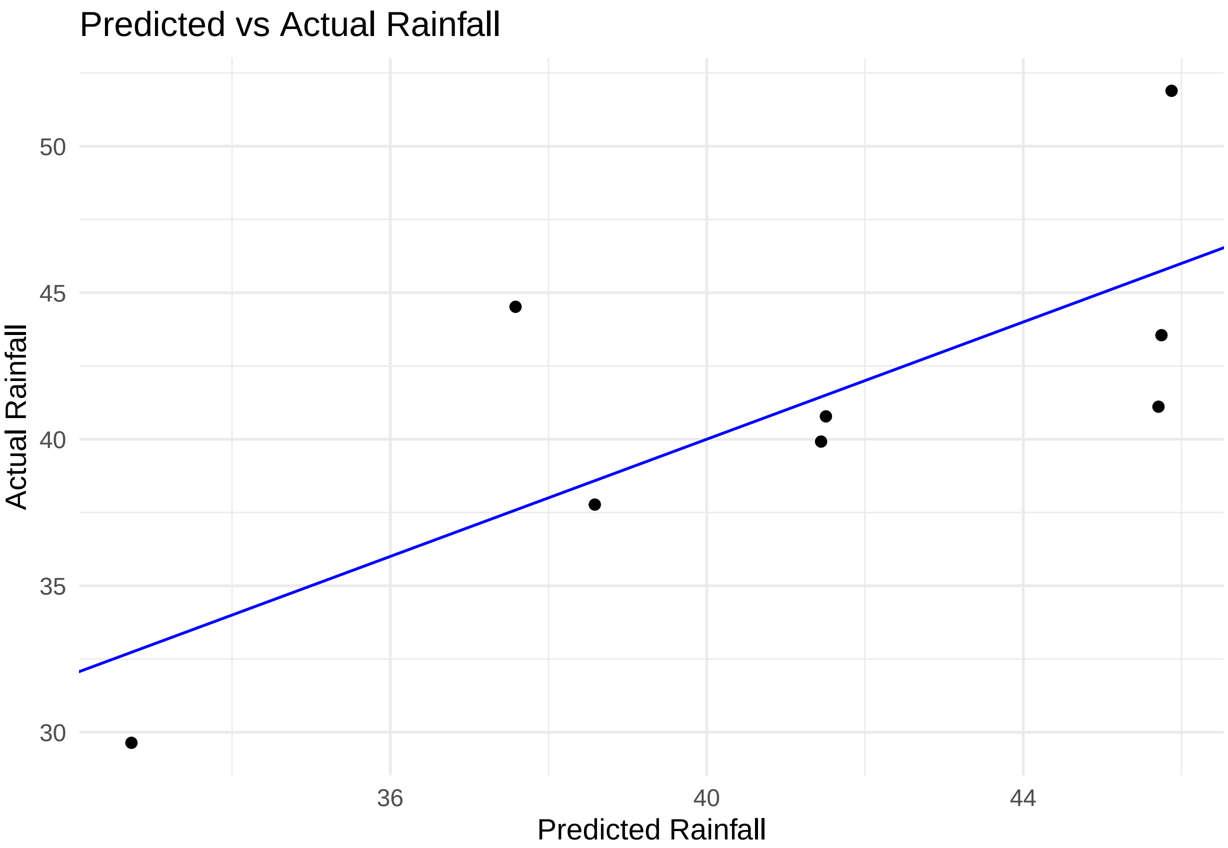
Building a simple model:

```r
# Building a linear regression model to predict total rainfall
rainfall_model <- lm(total_rainfall ~ avg_air_temp + avg_wind_speed, data = merged_data)

# Summarizing the model to see coefficients and significance
summary(rainfall_model)
```

```
##
## Call:
## lm(formula = total_rainfall ~ avg_air_temp + avg_wind_speed,
##     data = merged_data)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -0.8152  6.9367 -3.0886 -1.5250 -0.7262 -2.1973 -4.5994  6.0150
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       89.592     31.630   2.832   0.0366 *
## avg_air_temp       2.800      3.209   0.872   0.4229
## avg_wind_speed   -11.790      4.745  -2.485   0.0555 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.967 on 5 degrees of freedom
## Multiple R-squared:  0.5564, Adjusted R-squared:  0.3789
## F-statistic: 3.135 on 2 and 5 DF,  p-value: 0.1311
```

```r
# Visualizing the relationship between predicted and actual rainfall
predicted_rainfall <- predict(rainfall_model)

ggplot(merged_data, aes(x = predicted_rainfall, y = total_rainfall)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "blue") +
  labs(title = "Predicted vs Actual Rainfall", x = "Predicted Rainfall", y = "Actual Rainfall") +
  theme_minimal()
```



Used a linear regression model as it is easy to implement helps to understand the relationship between rainfall and the selected climate variables.

Predictor Variables: Average Air Temperature: We included this because it generally influences weather patterns and could impact rainfall. Average Wind Speed: Wind speed can also influence weather conditions.

Based on the scatter plot of Predicted vs Actual Rainfall, we can see that the model is making reasonable predictions, but there is still some predictions arent perfect