HW 4 Dplyr and tidyR

2024-09-24

a. Read in the data for all the years from 1985 to 2023.

```
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)

  header <- scan(path, what = 'character', nlines = 1)

  # After 2007, there is extra row for the header: the units of the first row (header).
  skip_lines <- ifelse(year >= 2007, 2, 1)
  buoy <- fread(path, header = FALSE, skip = skip_lines, fill = TRUE)

  # Check if the number of columns matches the header
  if (ncol(buoy) == length(header)) {
    colnames(buoy) <- header
  } else {
    if (ncol(buoy) > length(header)) {
      colnames(buoy) <- c(header, paste0("ExtraCol", seq(ncol(buoy) - length(header))))
    } else {
      colnames(buoy) <- header[1:ncol(buoy)]
    }
  }

  # Combine 'YYYY', 'YY', and '#YY' into a 'Year' column
  buoy <- buoy %>%
    mutate(
      YY = if ("YY" %in% colnames(buoy)) 1900 + as.numeric(YY) else NA_real_,
      Year = coalesce(
        if ("YYYY" %in% colnames(buoy)) as.numeric(YYYY) else NA_real_,
        if ("YY" %in% colnames(buoy)) YY else NA_real_,
        if ("#YY" %in% colnames(buoy)) as.numeric(`#YY`) else NA_real_
      )
    ) %>%
  select(-YY) %>%
  # Ensure YYYY is selected first
  select(Year, everything())

  # Create a proper datetime column using lubridate
  if (all(c("Year", "MM", "DD") %in% colnames(buoy))) {
    buoy <- buoy %>%
      mutate(
      Year = as.numeric(Year),
```

```
      MM = sprintf("%02d", as.numeric(MM)),
      DD = sprintf("%02d", as.numeric(DD))
      ) %>%
      mutate(datetime = ymd(paste0(Year, MM, DD)))
  }

  return(buoy)
}

years <- 1985:2023

all_buoy_data <- lapply(years, read_buoy_data)
```

```
## Warning in fread(path, header = FALSE, skip = skip_lines, fill = TRUE): Stopped
## early on line 5114. Expected 16 fields but found 17. Consider fill=TRUE and
## comment.char=. First discarded non-empty line: <<2000 08 01 00 78 4.3 5.1 0.58
## 8.33 5.36 999 1022.9 17.3 17.5 15.0 99.0 99.00>>
```

```
# Combine all data into a single data.table
combined_buoy_data <- rbindlist(all_buoy_data, fill = TRUE)

# Remove YYYY and #YY columns from the combine data table
combined_buoy_data <- combined_buoy_data %>%
  # First check if either "YYYY" or "#YY" exists, and handle accordingly
  mutate(
    YYYY = case_when(
      !is.na(as.numeric(YYYY)) ~ as.numeric(YYYY),  # Use existing YYYY if valid
      !is.na(as.numeric(`#YY`)) ~ as.numeric(`#YY`), # Use #YY if valid
      TRUE ~ NA_real_  # Otherwise, assign NA
    )
  ) %>%
  # Drop the YYYY and "#YY" column if it exists
  select(-YYYY, -`#YY`)
```

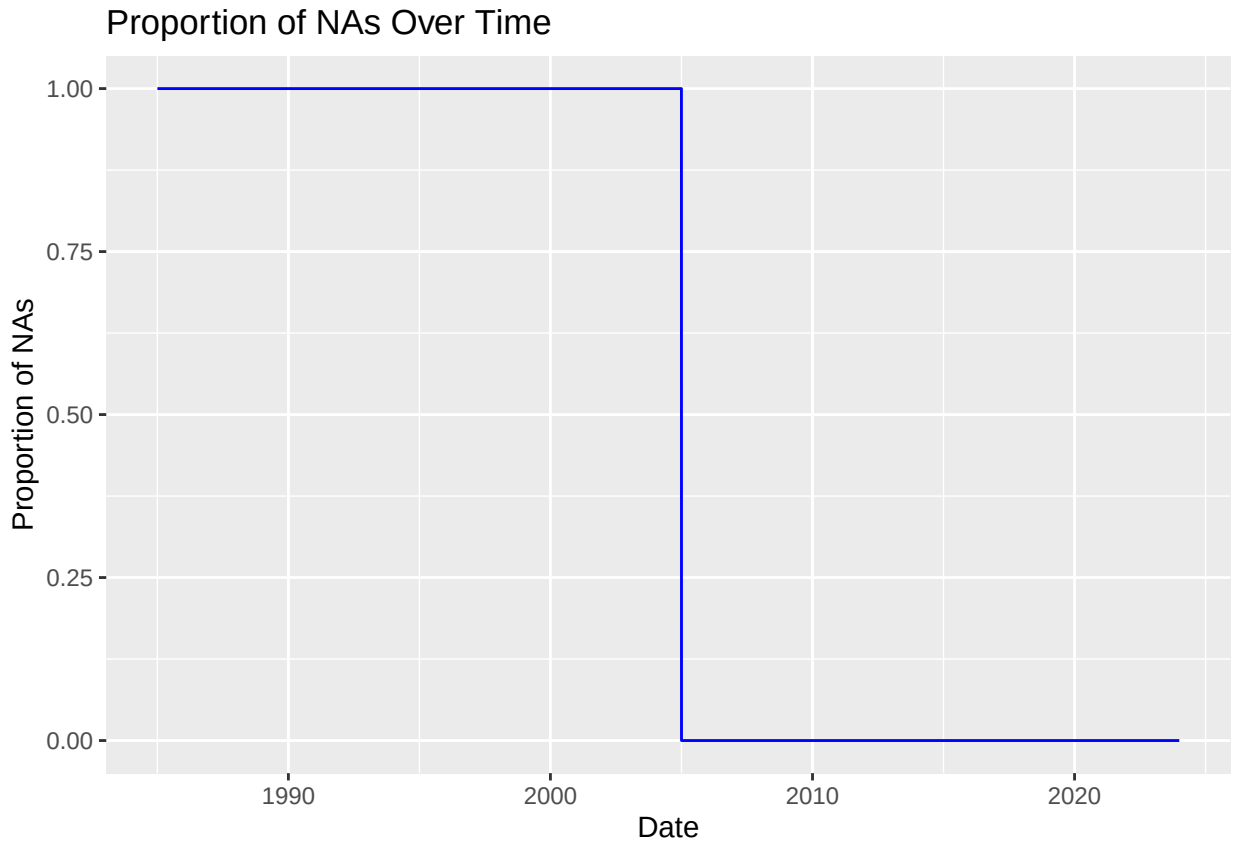
b. NA's

Convert the variables showed up as 999 to NA's in the dataset. It is not always appropriate to convert missing/null data to NA's. When the values 999 or 99 are actually meaningful in a specific context (for example, representing valid extreme values), replacing them with NA might distort the dataset and lead to incorrect conclusions.

```
combined_buoy_data <- combined_buoy_data %>%
  mutate(across(-datetime, ~ as.numeric(as.character(.)))) %>%  # Convert all columns to numeric except
  mutate(across(-datetime, ~ na_if(., 999))) %>%  # Replace 999 with NA, except for datetime
  mutate(across(-datetime, ~ na_if(., 99)))
  # As I remember in class, professor told that 99 is also consider a NA value.
```

Analyze the pattern of NA's The proportion of NAs is relatively high between 1985 to 2005, suggesting that data collection methods were less reliable because there are lack of information as compare with the information between 2005 to 2023. The proportion of NAs between 2005 and 2023 indicates that improvements in data collection methods or increased investment in buoy technology.

```r
# Create a summary of NAs as proportion of total observations
na_distribution <- combined_buoy_data %>%
  mutate(date = as.Date(datetime)) %>%  # Extract date
  group_by(date) %>%
  summarise(
    na_count = sum(is.na(mm)),  # Adjust 'mm' to your relevant column
    total_count = n(),           # Total number of observations
    na_proportion = na_count / total_count,  # Calculate proportion of NAs
    .groups = 'drop'
  ) %>%
  ungroup()

# Plotting NA distribution over time
ggplot(na_distribution, aes(x = date, y = na_proportion)) +
  geom_line(color = "blue") +
  labs(title = "Proportion of NAs Over Time", x = "Date", y = "Proportion of NAs")
```




```r
#combined_buoy_data <- combined_buoy_data %>%
#  left_join(shutdown_data, by = "date") %>%
#  left_join(budget_data, by = "date")
```

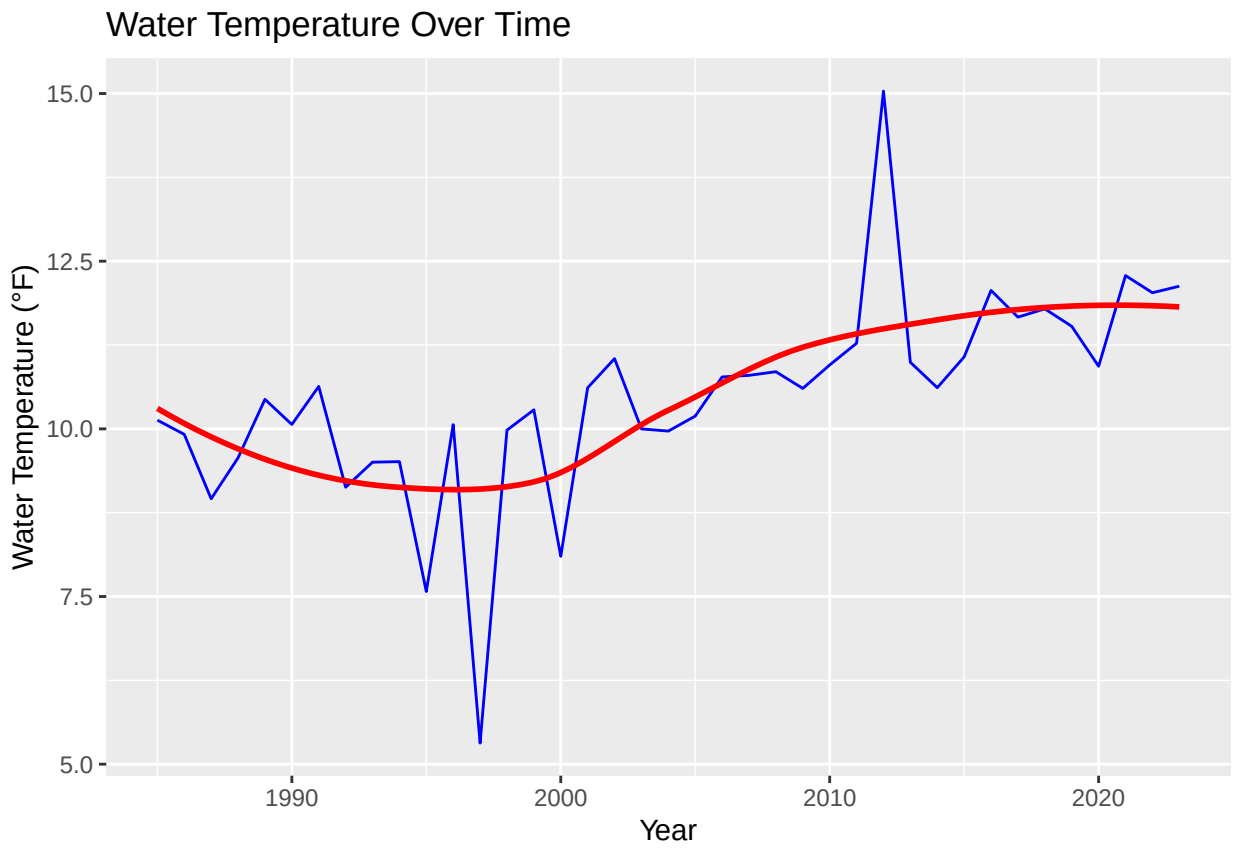
Bonus: Add other data sources that might shed light on the NA's

c. Use the Buoy data to see the effects of climate change.

Visualization The line plot shows how annual average water temperature has changed over time, with the smoothing line highlighting long-term trends. It shows the pattern of climate change from 1985 to 2023.

```
summary_data <- combined_buoy_data %>%
  group_by(Year) %>%
  summarise(
    avg_WTMP = mean(WTMP, na.rm = TRUE),  # Average WTMP per year
    .groups = 'drop'
  )

ggplot(summary_data, aes(x = Year, y = avg_WTMP)) +
  geom_line(color = "blue") +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(title = "Water Temperature Over Time",
       x = "Year", y = "Water Temperature (°F)")
```

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



Statistics Using the linear regression model, the estimated coefficient for the variable "Year" (slope) is approximately 0.08. This suggests that for each additional year, the average water temperature is expected to increase by 0.08 degrees Fahrenheit, holding all other variables constant. It is another way to show the effect of climate change from 1985 to 2023.

```r
model <- lm(avg_WTMP ~ Year, data = summary_data)
summary(model)
```

```
##
## Call:
## lm(formula = avg_WTMP ~ Year, data = summary_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5863 -0.2855  0.0819  0.3935  3.9124
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -152.88599   34.68132  -4.408 8.62e-05 ***
## Year           0.08152    0.01731   4.710 3.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.216 on 37 degrees of freedom
## Multiple R-squared:  0.3749, Adjusted R-squared:  0.358
## F-statistic: 22.19 on 1 and 37 DF,  p-value: 3.44e-05
```

d. Analyze the patterns between rainfall in Boston and weather buoy readings from 1985 to 2013

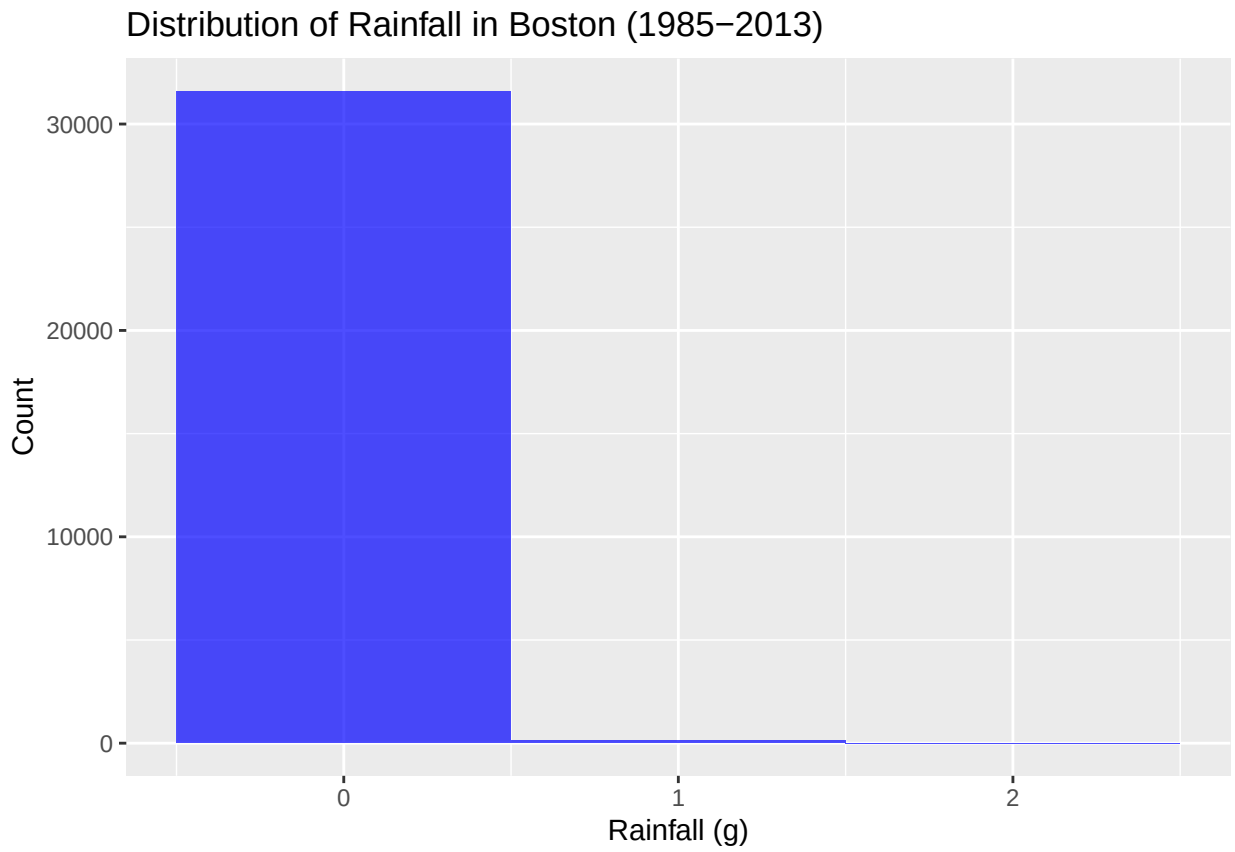
1) Data Acquisition As average surface temperatures on Earth increase, evaporation rates rise, leading to a subsequent increase in overall precipitation (rainfall).

```r
rainfall_data <- read.csv("Rainfall.csv")

# Summary statistics for rainfall
rainfall_summary <- rainfall_data %>%
  summarise(
    total_rainfall = sum(HPCP),
    mean_rainfall = mean(HPCP),
    median_rainfall = median(HPCP),
    count_days = n()
  )

# Summary statistics for buoy data
buoy_summary <- combined_buoy_data %>%
  summarise(
    mean_temp = mean(WTMP, na.rm = TRUE),
    median_temp = median(WTMP, na.rm = TRUE),
    count_days = n()
  )
```

Summary Statistics:

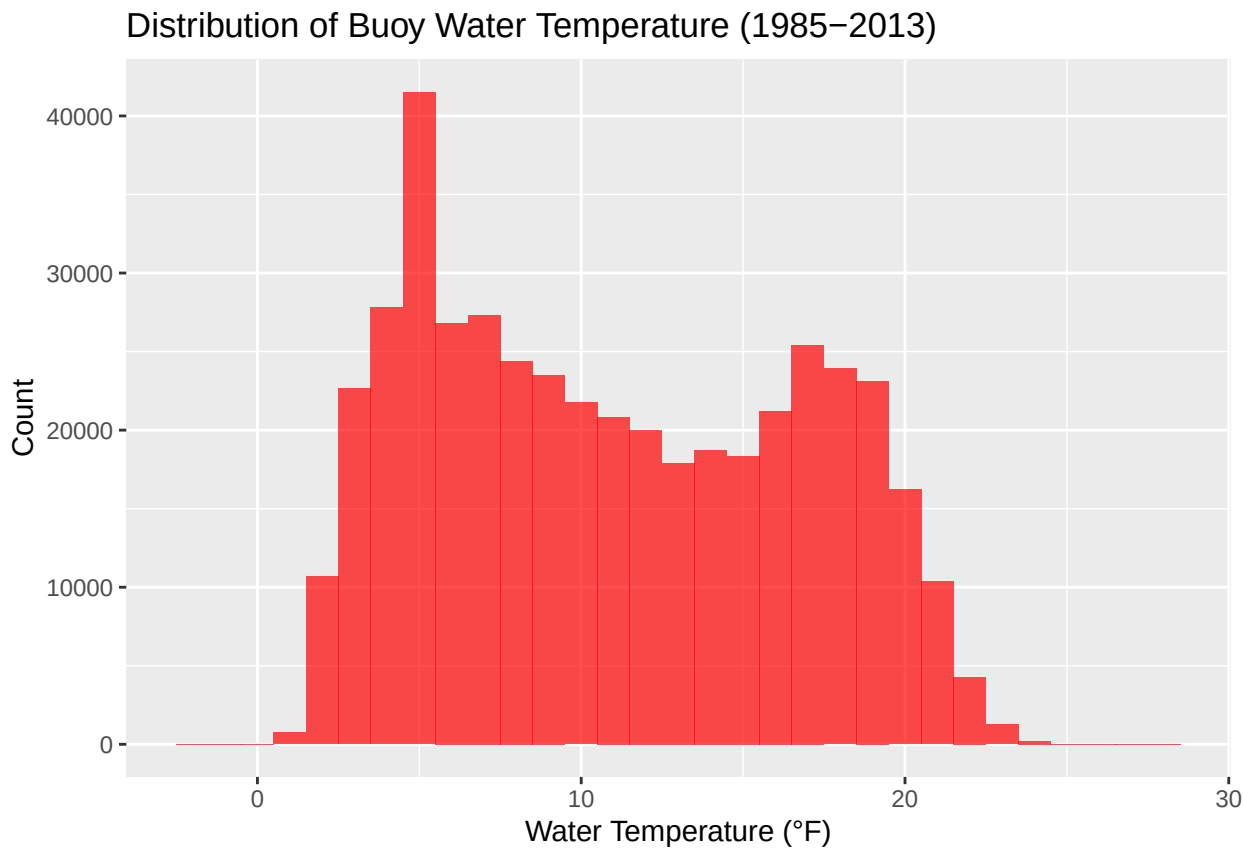
Visualization Distributions \\

```
# Histogram of total yearly rainfall
ggplot(rainfall_data, aes(x = HPCP)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Rainfall in Boston (1985-2013)",
       x = "Rainfall (g)", y = "Count")
```




```
# Histogram of buoy temperature
ggplot(combined_buoy_data, aes(x = WTMP)) +
  geom_histogram(binwidth = 1, fill = "red", alpha = 0.7) +
  labs(title = "Distribution of Buoy Water Temperature (1985-2013)",
       x = "Water Temperature (°F)", y = "Count")
```

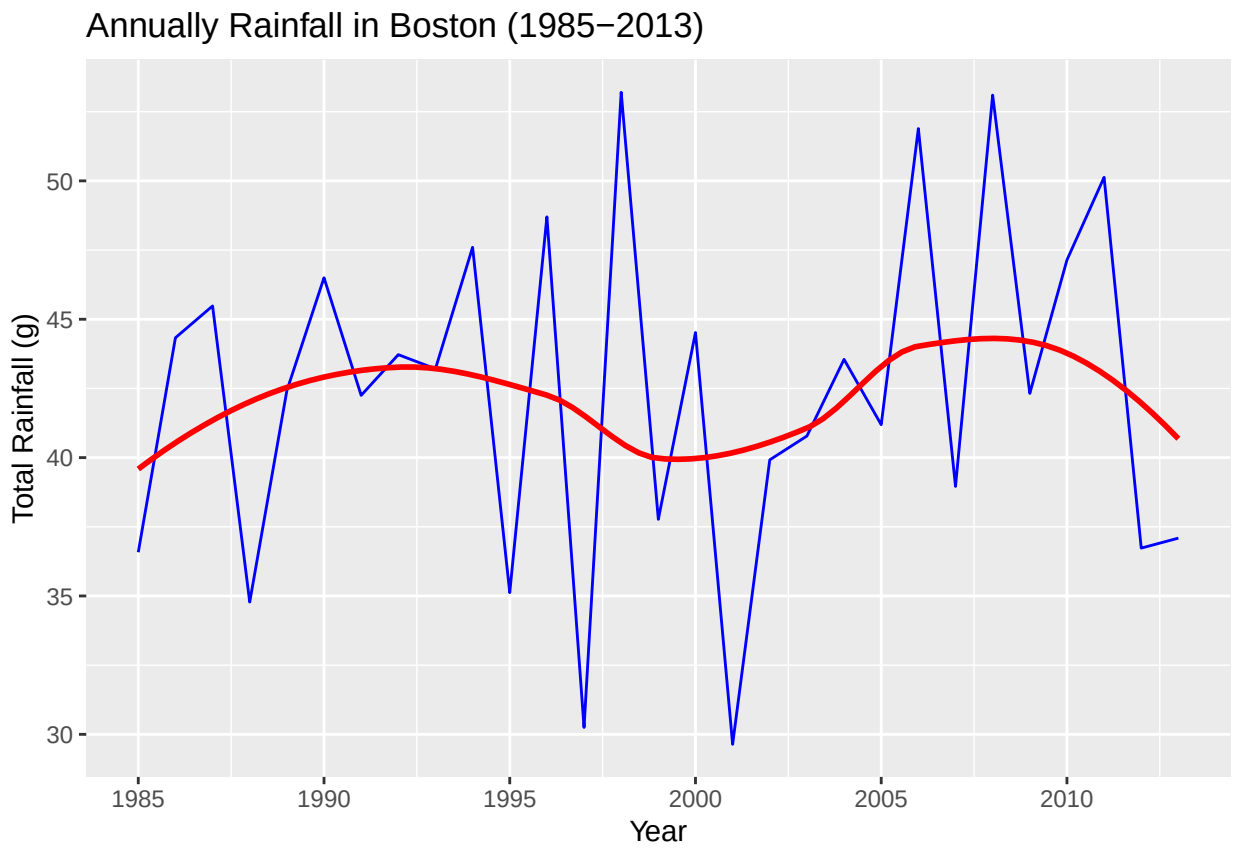
```
## Warning: Removed 13186 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



2) Visualizations. The smoothing line in the visualization reveals a similar wave pattern between rainfall and buoy readings, indicating a potential relationship between these two variables.

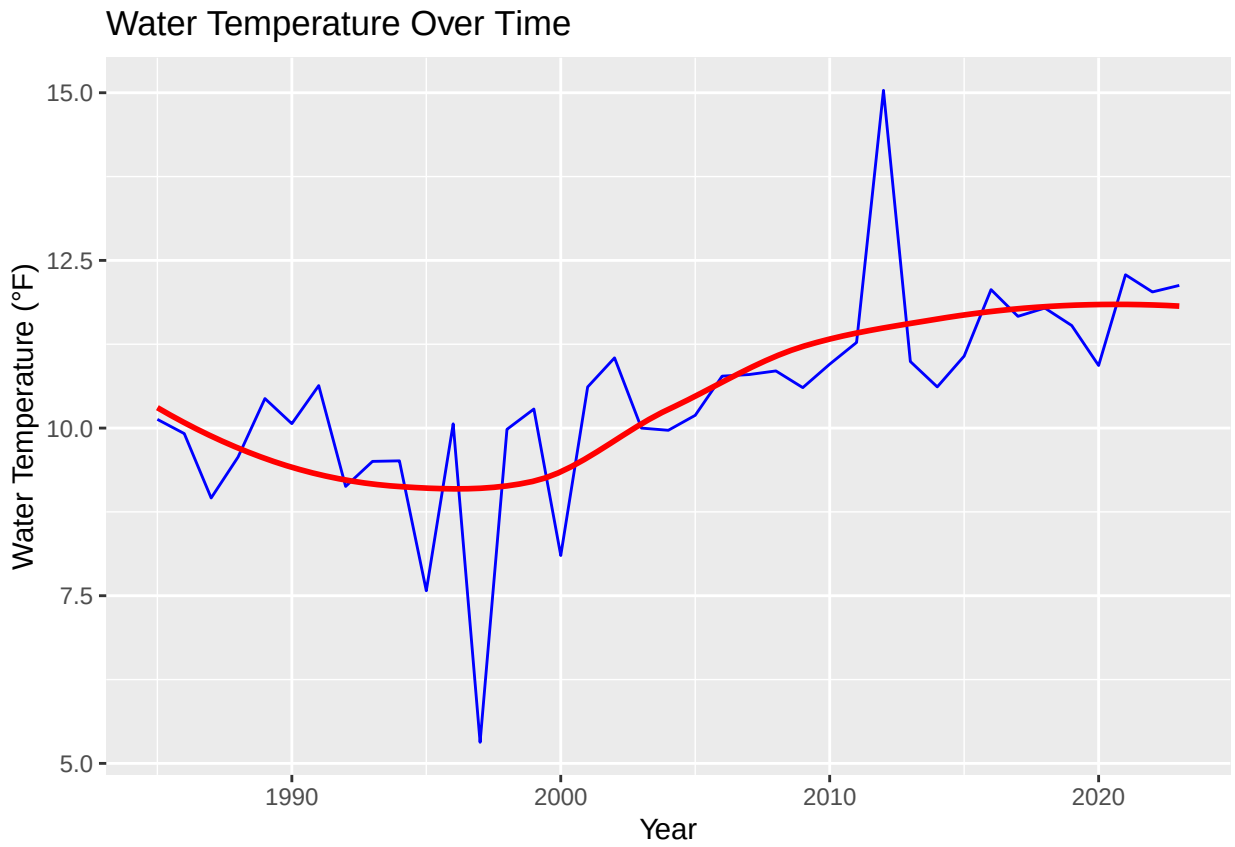
```
# Visualization pattern for rainfall
rainfall_yearly <- rainfall_data %>%
  mutate(year = str_sub(DATE, 1, 4)) %>%
  group_by(year) %>%
  summarise(total_rainfall = sum(HPCP), .groups = 'drop')

ggplot(rainfall_yearly, aes(x = as.numeric(year), y = total_rainfall)) +
  geom_line(color = "blue") +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(title = "Annually Rainfall in Boston (1985-2013)",
       x = "Year", y = "Total Rainfall (g)")
```

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




```
# Visualization pattern for buoy
summary_data <- combined_buoy_data %>%
  group_by(Year) %>%
  summarise(avg_WTMP = mean(WTMP, na.rm = TRUE), .groups = 'drop')

ggplot(summary_data, aes(x = Year, y = avg_WTMP)) +
  geom_line(color = "blue") +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(title = "Water Temperature Over Time",
       x = "Year", y = "Water Temperature (°F)")
```

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



3) Try building a very simple model. Using the linear regression model for both rainfall and buoy, the estimated coefficient for the variable "year" (slope) in rainfall is approximately 0.07, and in buoy is approximately 0.08. It seems very close to each other.

The slope in rainfall suggests that for each additional year, the total rainfall is expected to increase by 0.07 g, holding all other variables constant. The slope in buoy suggests that for each additional year, the average water temperature is expected to increase by 0.08 degrees Fahrenheit, holding all other variables constant.

It suggests a consistent trend of increasing rainfall and water temperature over the years, which can be indicative of broader climatic changes. This analysis opens avenues for deeper exploration into the relationship between these two variables and their implications for environmental policy, resource management, and further research into climate change impacts.

After going through the process of analyzing rainfall and buoy data underscores the difficulties in making accurate predictions about weather patterns. It fosters a greater appreciation for the expertise meteorologists bring to the table and the continuous advancements in meteorological science. Weather forecasting is a blend of art and science, and understanding this nuance can indeed lead to more sympathy for weather professionals, recognizing the hard work and dedication they put into their forecasts despite the inherent uncertainties.

```
# Regression model for rainfall
model <- lm(total_rainfall ~ as.numeric(year), data = rainfall_yearly)
summary(model)
```

```
##
## Call:
## lm(formula = total_rainfall ~ as.numeric(year), data = rainfall_yearly)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8761  -4.6048   0.7516   3.9881  10.8958
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -98.83532  277.66711  -0.356    0.725
## as.numeric(year)   0.07064    0.13890   0.509    0.615
##
## Residual standard error: 6.258 on 27 degrees of freedom
## Multiple R-squared:  0.009488,   Adjusted R-squared:  -0.0272
## F-statistic: 0.2586 on 1 and 27 DF,  p-value: 0.6152
```

```r
# Regression model for buoy
model <- lm(avg_WTMP ~ Year, data = summary_data)
summary(model)
```

```
##
## Call:
## lm(formula = avg_WTMP ~ Year, data = summary_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5863 -0.2855  0.0819  0.3935  3.9124
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -152.88599   34.68132  -4.408 8.62e-05 ***
## Year           0.08152    0.01731   4.710 3.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.216 on 37 degrees of freedom
## Multiple R-squared:  0.3749, Adjusted R-squared:  0.358
## F-statistic: 22.19 on 1 and 37 DF,  p-value: 3.44e-05
```