HW4

Dayu Tie

2024-09-23

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
library(lubridate)
```

```
data=read.csv('Rainfall.csv')
head(data)
```

```
##       STATION                                STATION_NAME           DATE HPCP
## 1 COOP:190770 BOSTON LOGAN INTERNATIONAL AIRPORT MA US 19850101 01:00 0.00
## 2 COOP:190770 BOSTON LOGAN INTERNATIONAL AIRPORT MA US 19850101 09:00 0.01
## 3 COOP:190770 BOSTON LOGAN INTERNATIONAL AIRPORT MA US 19850101 10:00 0.01
## 4 COOP:190770 BOSTON LOGAN INTERNATIONAL AIRPORT MA US 19850101 11:00 0.01
## 5 COOP:190770 BOSTON LOGAN INTERNATIONAL AIRPORT MA US 19850101 12:00 0.01
## 6 COOP:190770 BOSTON LOGAN INTERNATIONAL AIRPORT MA US 19850101 13:00 0.01
##   Measurement.Flag Quality.Flag
## 1                g           NA
## 2                            NA
## 3                            NA
## 4                            NA
## 5                            NA
## 6                            NA
```

```
library(data.table)
```

```
##
## 载入程序包：'data.table'
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(lubridate)

years <- 1985:2023
file_root <- "https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
tail <- ".txt.gz&dir=data/historical/stdmet/"
buoy_data_list <- list()

for (year in years) {

  if (year == 2000) next  # Skip the year 2000 as per your requirement

  path <- paste0(file_root, year, tail)
  skip_lines <- if (year >= 1985 & year <= 2006) 1 else 2

  header <- scan(path, what = 'character', nlines = 1, quiet = TRUE)
  second_line <- scan(path, what = 'character', nlines = 1, skip = 1, quiet = TRUE)

  if (length(second_line) == length(header)) {
    buoy <- fread(path, header = FALSE, skip = skip_lines)
  } else {
    buoy <- fread(path, header = FALSE, skip = 1)
  }

  colnames(buoy) <- header


  if (all(c('YY', 'MM', 'DD', 'hh', 'mm') %in% colnames(buoy))) {
    buoy[, Date := make_datetime(YY, MM, DD, hh, mm)]
  }

  buoy_data_list[[as.character(year)]] <- buoy
}
all_buoy_data <- rbindlist(buoy_data_list, fill = TRUE)
head(all_buoy_data)
```

```
##       YY    MM    DD    hh    WD  WSPD   GST  WVHT   DPD   APD   MWD    BAR
##    <int> <int> <int> <int> <int> <num> <num> <num> <num> <num> <int>  <num>
## 1:    85     1     1     0    60     4     5    99    99    99   999 1030.3
## 2:    85     1     1     1    80     4     5    99    99    99   999 1030.0
## 3:    85     1     1     2   100     4     5    99    99    99   999 1030.1
## 4:    85     1     1     3   100     4     5    99    99    99   999 1029.4
## 5:    85     1     1     4   110     4     5    99    99    99   999 1028.6
## 6:    85     1     1     5    90     4     5    99    99    99   999 1027.8
##     ATMP  WTMP  DEWP   VIS  YYYY  TIDE    mm   #YY  WDIR  PRES
##    <num> <num> <num> <num> <int> <num> <int> <int> <int> <num>
## 1:   4.7   6.7   999    99    NA    NA    NA    NA    NA    NA
## 2:   5.1   6.7   999    99    NA    NA    NA    NA    NA    NA
## 3:   5.6   6.6   999    99    NA    NA    NA    NA    NA    NA
## 4:   5.8   6.7   999    99    NA    NA    NA    NA    NA    NA
## 5:   5.8   6.7   999    99    NA    NA    NA    NA    NA    NA
## 6:   5.3   6.7   999    99    NA    NA    NA    NA    NA    NA
```

```
all_buoy_data$YY <- as.character(all_buoy_data$YY)
all_buoy_data$YYYY <- as.character(all_buoy_data$YYYY)
all_buoy_data$`#YY` <- as.character(all_buoy_data$`#YY`)
all_buoy_data$year <- ifelse(!is.na(all_buoy_data$YYYY), all_buoy_data$YYYY,
                             ifelse(!is.na(all_buoy_data$`#YY`), all_buoy_data$`#YY`, all_buoy_dat
a$YY))
all_buoy_data$year <- as.numeric(all_buoy_data$year)
all_buoy_data$year <- ifelse(all_buoy_data$year >= 85 & all_buoy_data$year <= 98,
                             all_buoy_data$year + 1900,
                             all_buoy_data$year)
all_buoy_data <- all_buoy_data %>% select(-YY, -YYYY, -`#YY`)
head(all_buoy_data)
```

```
##       MM    DD    hh    WD  WSPD   GST  WVHT   DPD   APD   MWD    BAR  ATMP
##    <int> <int> <int> <int> <num> <num> <num> <num> <num> <int>  <num> <num>
## 1:     1     1     0    60     4     5    99    99    99   999 1030.3   4.7
## 2:     1     1     1    80     4     5    99    99    99   999 1030.0   5.1
## 3:     1     1     2   100     4     5    99    99    99   999 1030.1   5.6
## 4:     1     1     3   100     4     5    99    99    99   999 1029.4   5.8
## 5:     1     1     4   110     4     5    99    99    99   999 1028.6   5.8
## 6:     1     1     5    90     4     5    99    99    99   999 1027.8   5.3
##     WTMP  DEWP   VIS  TIDE    mm  WDIR  PRES  year
##    <num> <num> <num> <num> <int> <int> <num> <num>
## 1:   6.7   999    99    NA    NA    NA    NA  1985
## 2:   6.7   999    99    NA    NA    NA    NA  1985
## 3:   6.6   999    99    NA    NA    NA    NA  1985
## 4:   6.7   999    99    NA    NA    NA    NA  1985
## 5:   6.7   999    99    NA    NA    NA    NA  1985
## 6:   6.7   999    99    NA    NA    NA    NA  1985
```

```
all_buoy_data <- all_buoy_data %>% relocate(year, .before = 1)
all_buoy_data$PRES <- ifelse(!is.na(all_buoy_data$PRES), all_buoy_data$PRES, all_buoy_data$BAR)
all_buoy_data$WDIR <- ifelse(!is.na(all_buoy_data$WDIR), all_buoy_data$WDIR, all_buoy_data$WD)
all_buoy_data <- all_buoy_data %>% select(-BAR, -WD)
head(all_buoy_data)
```

```
##     year    MM    DD    hh  WSPD   GST  WVHT   DPD   APD   MWD  ATMP  WTMP
##    <num> <int> <int> <int> <num> <num> <num> <num> <num> <int> <num> <num>
## 1:  1985     1     1     0     4     5    99    99    99   999   4.7   6.7
## 2:  1985     1     1     1     4     5    99    99    99   999   5.1   6.7
## 3:  1985     1     1     2     4     5    99    99    99   999   5.6   6.6
## 4:  1985     1     1     3     4     5    99    99    99   999   5.8   6.7
## 5:  1985     1     1     4     4     5    99    99    99   999   5.8   6.7
## 6:  1985     1     1     5     4     5    99    99    99   999   5.3   6.7
##     DEWP   VIS  TIDE    mm  WDIR   PRES
##    <num> <num> <num> <int> <int>  <num>
## 1:   999    99    NA    NA    60 1030.3
## 2:   999    99    NA    NA    80 1030.0
## 3:   999    99    NA    NA   100 1030.1
## 4:   999    99    NA    NA   100 1029.4
## 5:   999    99    NA    NA   110 1028.6
## 6:   999    99    NA    NA    90 1027.8
```

```
library(data.table)
file_root <- "https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
year <- "2000"
tail <- ".txt.gz&dir=data/historical/stdmet/"
path <- paste0(file_root, year, tail)
buoy_2000 <- fread(path, header = FALSE, skip = 1, fill = TRUE)
```

```
## Warning in fread(path, header = FALSE, skip = 1, fill = TRUE): Stopped early on
## line 5114. Expected 16 fields but found 17. Consider fill=17 or even more based
## on your knowledge of the input file. Use fill=Inf for reading the whole file
## for detecting the number of fields. 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>>
```

```
header <- scan(path, what = 'character', nlines = 1, quiet = TRUE)
if (length(header) != ncol(buoy_2000)) {
  header <- header[1:ncol(buoy_2000)]
}
setnames(buoy_2000, header)
if (!"TIDE" %in% colnames(buoy_2000)) {
  buoy_2000[, TIDE := 99]
} else {
  buoy_2000[TIDE == "", TIDE := 99]
}
buoy_2000 <- buoy_2000 %>% rename(year = YYYY)
buoy_2000 <- buoy_2000 %>%
  rename(PRES = BAR, WDIR = WD)
buoy_2000$mm <- NA
buoy_2000$year <- as.numeric(buoy_2000$year)
head(buoy_2000)
```

```
##     year    MM    DD    hh  WDIR  WSPD   GST  WVHT   DPD   APD   MWD   PRES
##    <num> <int> <int> <int> <int> <num> <num> <num> <num> <num> <int>  <num>
## 1:  2000     1     1     0   315   0.8   1.5  0.54 10.00  4.55   999 1019.2
## 2:  2000     1     1     1   271   0.7   1.9  0.53  4.55  4.61   999 1020.3
## 3:  2000     1     1     2   232   2.3   3.1  0.52  4.76  4.78   999 1020.4
## 4:  2000     1     1     3   236   2.9   3.8  0.52 10.00  4.86   999 1021.0
## 5:  2000     1     1     4   232   4.1   4.9  0.50 10.00  5.00   999 1021.2
## 6:  2000     1     1     5   228   5.4   6.5  0.46  4.55  4.94   999 1021.5
##     ATMP  WTMP  DEWP   VIS  TIDE     mm
##    <num> <num> <num> <num> <num> <lgcl>
## 1:   1.1   5.9  -3.6    99    99     NA
## 2:   1.2   5.9  -3.6    99    99     NA
## 3:   1.5   5.8  -3.3    99    99     NA
## 4:   1.3   5.9  -3.6    99    99     NA
## 5:   0.9   5.9  -4.0    99    99     NA
## 6:   0.7   5.9  -6.5    99    99     NA
```

```
buoy_1985_2023 <- bind_rows(buoy_2000, all_buoy_data)
buoy_1985_2023 <- buoy_1985_2023 %>% arrange(year)
buoy_1985_2023$mm[is.na(buoy_1985_2023$mm)] <- 0
head(buoy_1985_2023)
```

```
##     year    MM    DD    hh  WDIR  WSPD   GST  WVHT   DPD   APD   MWD   PRES
##    <num> <int> <int> <int> <int> <num> <num> <num> <num> <num> <int>  <num>
## 1:  1985     1     1     0    60     4     5    99    99    99   999 1030.3
## 2:  1985     1     1     1    80     4     5    99    99    99   999 1030.0
## 3:  1985     1     1     2   100     4     5    99    99    99   999 1030.1
## 4:  1985     1     1     3   100     4     5    99    99    99   999 1029.4
## 5:  1985     1     1     4   110     4     5    99    99    99   999 1028.6
## 6:  1985     1     1     5    90     4     5    99    99    99   999 1027.8
##     ATMP  WTMP  DEWP   VIS  TIDE    mm
##    <num> <num> <num> <num> <num> <num>
## 1:   4.7   6.7   999    99    NA     0
## 2:   5.1   6.7   999    99    NA     0
## 3:   5.6   6.6   999    99    NA     0
## 4:   5.8   6.7   999    99    NA     0
## 5:   5.8   6.7   999    99    NA     0
## 6:   5.3   6.7   999    99    NA     0
```

```
buoy_1985_2023$datetime <- make_datetime(
  year = buoy_1985_2023$year,
  month = buoy_1985_2023$MM,
  day = buoy_1985_2023$DD,
  hour = buoy_1985_2023$hh,
  min = buoy_1985_2023$mm
)
```

##(b)

```
buoy_1985_2023[buoy_1985_2023 == 999] <- NA
colSums(is.na(buoy_1985_2023))
```

```
##     year       MM       DD       hh     WDIR     WSPD      GST     WVHT
##        0        0        0        0    43556        0        0        0
##      DPD      APD      MWD     PRES     ATMP     WTMP     DEWP      VIS
##        0        0   325297      261   102761    13186   253613        0
##     TIDE       mm datetime
##   124498        0        0
```

##It’s not always appropriate to convert missing/null data to NA.If a placeholder value (like 999, -1, or similar) conveys a specific meaning, such as ‘data not collected’ or ‘not applicable,’ it’s best to keep it as is. Changing it to NA might result in losing valuable context.NA bascily exist in WDIR and TIDE after 2000 because TIDE is a new factor .

```
library(openxlsx)
write.xlsx(buoy_1985_2023, "buoy_1985_2023.xlsx")
```

##(c)

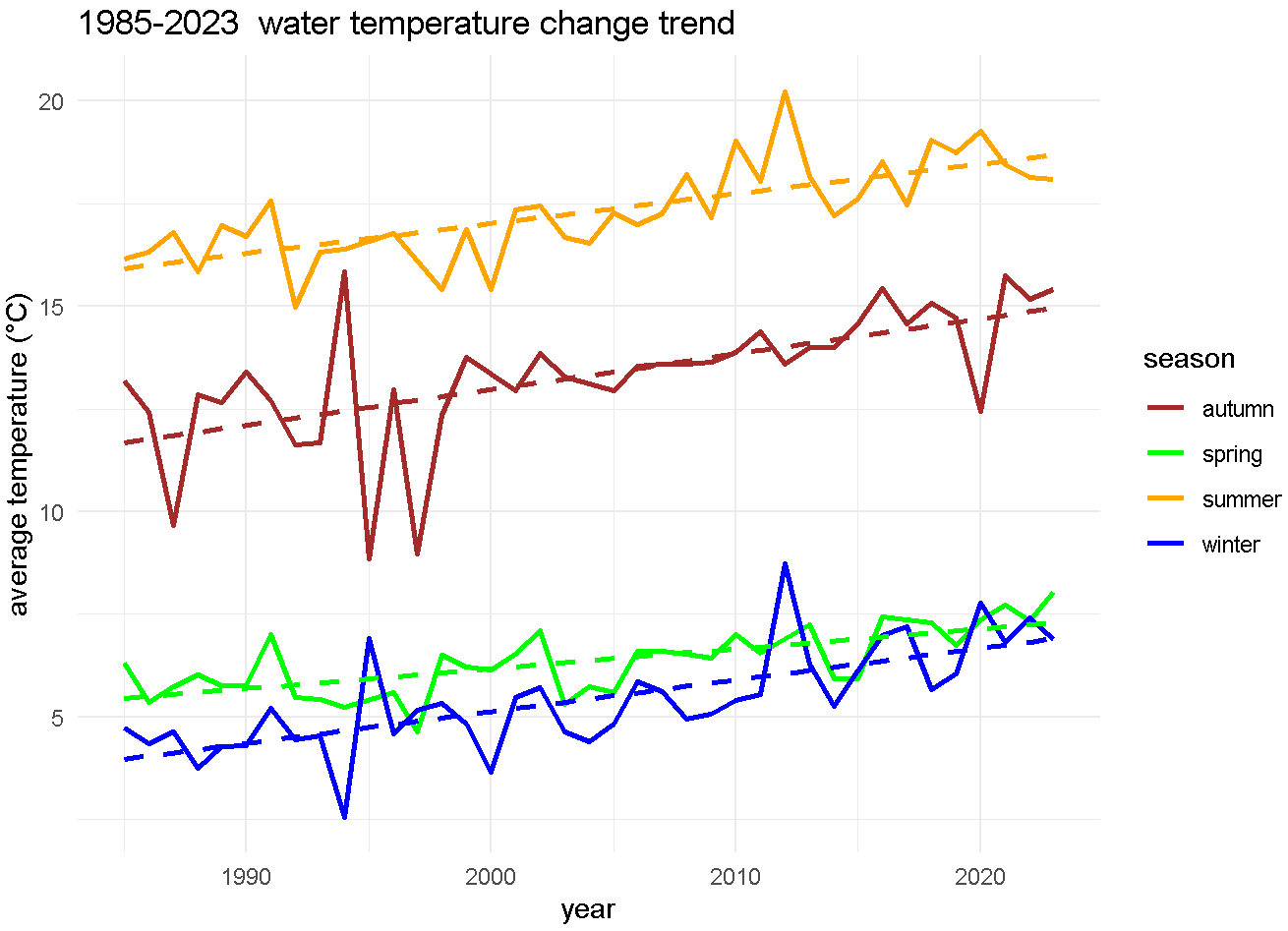
```
library(readxl)
library(ggplot2)
library(dplyr)
library(lubridate)


data <- read_excel("buoy_1985_2023.xlsx")
data$datetime <- as.Date(data$datetime)
temperature_data <- data %>% filter(!is.na(WTMP))
temperature_data$season <- case_when(
  month(temperature_data$datetime) %in% c(3, 4, 5) ~ "spring",
  month(temperature_data$datetime) %in% c(6, 7, 8) ~ "summer",
  month(temperature_data$datetime) %in% c(9, 10, 11) ~ "autumn",
  TRUE ~ "winter"  # 12, 1, 2
)

# every season's average temperature
seasonal_avg_temp <- temperature_data %>%
  group_by(season, year = year(datetime)) %>%
  summarise(avg_wtmp = mean(WTMP, na.rm = TRUE)) %>%
  ungroup()
```

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

```
# add trend line
ggplot(seasonal_avg_temp, aes(x = year, y = avg_wtmp, color = season)) +
  geom_line(size = 1) +
  geom_smooth(method = "lm", aes(group = season), se = FALSE, linetype = "dashed") +
  labs(title = "1985-2023  water temperature change trend", x = "year", y = "average temperature
(°C)") +
  theme_minimal() +
  scale_color_manual(values = c("spring" = "green", "summer" = "orange", "autumn" = "brown", "wint
er" = "blue"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



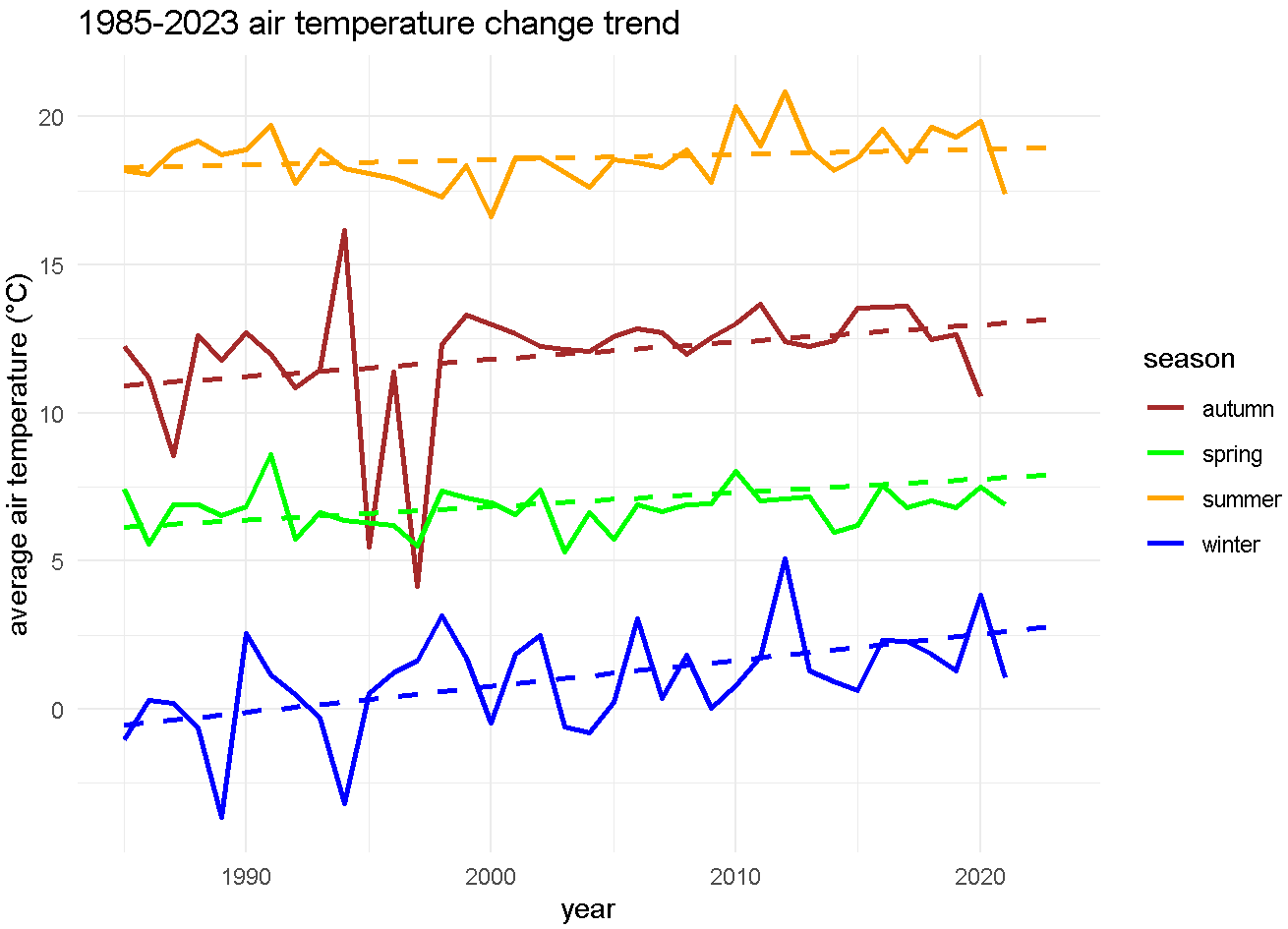

```
seasonal_avg_temp <- temperature_data %>%
  group_by(season, year = year(datetime)) %>%
  summarise(avg_atmp = mean(ATMP, na.rm = TRUE)) %>%
  ungroup()
```

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

```
# air temperature and trend line
ggplot(seasonal_avg_temp, aes(x = year, y = avg_atmp, color = season)) +
  geom_line(size = 1) +
  geom_smooth(method = "lm", aes(group = season), se = FALSE, linetype = "dashed") +
  labs(title = "1985-2023 air temperature change trend ", x = "year", y = "average air temperature
(°C)") +
  theme_minimal() +
  scale_color_manual(values = c("spring" = "green", "summer" = "orange", "autumn" = "brown", "wint
er" = "blue"))
```

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

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```




```
correlation_data <- data %>%
  filter(!is.na(ATMP) & !is.na(WTMP))
correlation <- cor(correlation_data$ATMP, correlation_data$WTMP)
correlation
```

```
## [1] 0.8880785
```

##As global temperatures rise persistently, the heat from the air is directly transferred to water bodies, resulting in an increase in water temperature. This effect is particularly pronounced in coastal regions and still water bodies like lakes.Positive correlation coefficient 0.8880785 indicates a strong correlation between water and air temperatures, which can be used to validate the accuracy of climate models or to assess the extent to which different regions are responding to global warming.

##(d)

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

buoy_data <- read_excel("buoy_1985_2023.xlsx")
rainfall_data <- read.csv("Rainfall.csv")



rainfall_data$DATE <- as.POSIXct(rainfall_data$DATE, format="%Y%m%d %H:%M")

rainfall_data$date_only <- as.Date(rainfall_data$DATE)
buoy_data$datetime <- as.POSIXct(paste(buoy_data$year, buoy_data$MM, buoy_data$DD, buoy_data$hh, s
ep="-"), format="%Y-%m-%d-%H")
buoy_data$date_only <- as.Date(buoy_data$datetime)

merged_data <- inner_join(buoy_data, rainfall_data, by = "date_only")
```

```
## Warning in inner_join(buoy_data, rainfall_data, by = "date_only"): Detected an unexpected many-
to-many relationship between `x` and `y`.
## ℹ Row 1 of `x` matches multiple rows in `y`.
## ℹ Row 1 of `y` matches multiple rows in `x`.
## ℹ If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
merged_data
```

```
## # A tibble: 711,356 × 26
##     year    MM    DD    hh  WDIR  WSPD   GST  WVHT   DPD   APD MWD    PRES  ATMP
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl> <dbl> <dbl>
##  1  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  2  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  3  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  4  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  5  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  6  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  7  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  8  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
##  9  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
## 10  1985     1     1     0    60     4     5    99    99    99 NA    1030.   4.7
## # ℹ 711,346 more rows
## # ℹ 13 more variables: WTMP <dbl>, DEWP <lgl>, VIS <dbl>, TIDE <lgl>, mm <dbl>,
## #   datetime <dttm>, date_only <date>, STATION <chr>, STATION_NAME <chr>,
## #   DATE <dttm>, HPCP <dbl>, Measurement.Flag <chr>, Quality.Flag <lgl>
```

```
merged_data$WDIR[(merged_data$WDIR)==0] <- 1
model <- lm(HPCP ~ log(WDIR), data = merged_data)
summary(model)
```

```
##
## Call:
## lm(formula = HPCP ~ log(WDIR), data = merged_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04575 -0.03739 -0.02743  0.00264  1.99329
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.575e-02  4.168e-04  109.77   <2e-16 ***
## log(WDIR)   -1.544e-03  8.742e-05  -17.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07629 on 669928 degrees of freedom
##   (因为不存在，41426个观察量被删除了)
## Multiple R-squared:  0.0004651,  Adjusted R-squared:  0.0004636
## F-statistic: 311.7 on 1 and 669928 DF,  p-value: < 2.2e-16
```