HW 4

Zilu Sun

Fall 2024

(a)

```
library(data.table)
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 <- tryCatch(scan(path, what = 'character', nlines = 1), error = function(e) NULL)
  if (is.null(header)) return(NULL)
  skip_value <- ifelse(year < 2007, 1, 2)
  buoy <- fread(path, header = FALSE, skip = skip_value, fill = TRUE)
  buoy[, Year := year]
  if (year == 2000) {
    buoy <- cbind(buoy, NA)
    header <- c(header, "new_column")
  }
  if (ncol(buoy) < length(header)) {
    missing_cols <- length(header) - ncol(buoy)
    buoy <- cbind(buoy, matrix(NA, nrow = nrow(buoy), ncol = missing_cols))
  }
  colnames(buoy) <- c(header, "Year")[1:ncol(buoy)]

  return(buoy)
}
years <- 1985:2023
buoy_data_list <- lapply(years, read_buoy_data)
```

```
## Warning in fread(path, header = FALSE, skip = skip_value, 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>>
```

```
all_buoy_data_2 <- rbindlist(buoy_data_list, use.names = TRUE, fill = TRUE)
```

(b)

```r
missing_summary_999 <- sapply(all_buoy_data_2, function(x) sum(x == 999, na.rm = TRUE))
print(missing_summary_999)
```

```
##         YY         MM         DD         hh         WD       WSPD        GST
##          0          0          0          0      15290          0          0
##       WVHT        DPD        APD        MWD        BAR       ATMP       WTMP
##          0          0          0     325297         87     102761      13186
##       DEWP        VIS       Year       YYYY       TIDE new_column         mm
##     253613          0          0          0          0          0          0
##        #YY       WDIR       PRES
##          0      28266        174
```

```r
missing_summary_99 <- sapply(all_buoy_data_2, function(x) sum(x == 99, na.rm = TRUE))
print(missing_summary_99)
```

```
##         YY         MM         DD         hh         WD       WSPD        GST
##          0          0          0          0        232      33183      33485
##       WVHT        DPD        APD        MWD        BAR       ATMP       WTMP
##     144269     147961     144269       1870          0          0          0
##       DEWP        VIS       Year       YYYY       TIDE new_column         mm
##          0     443062          0          0     332691          0          0
##        #YY       WDIR       PRES
##          0        387          0
```

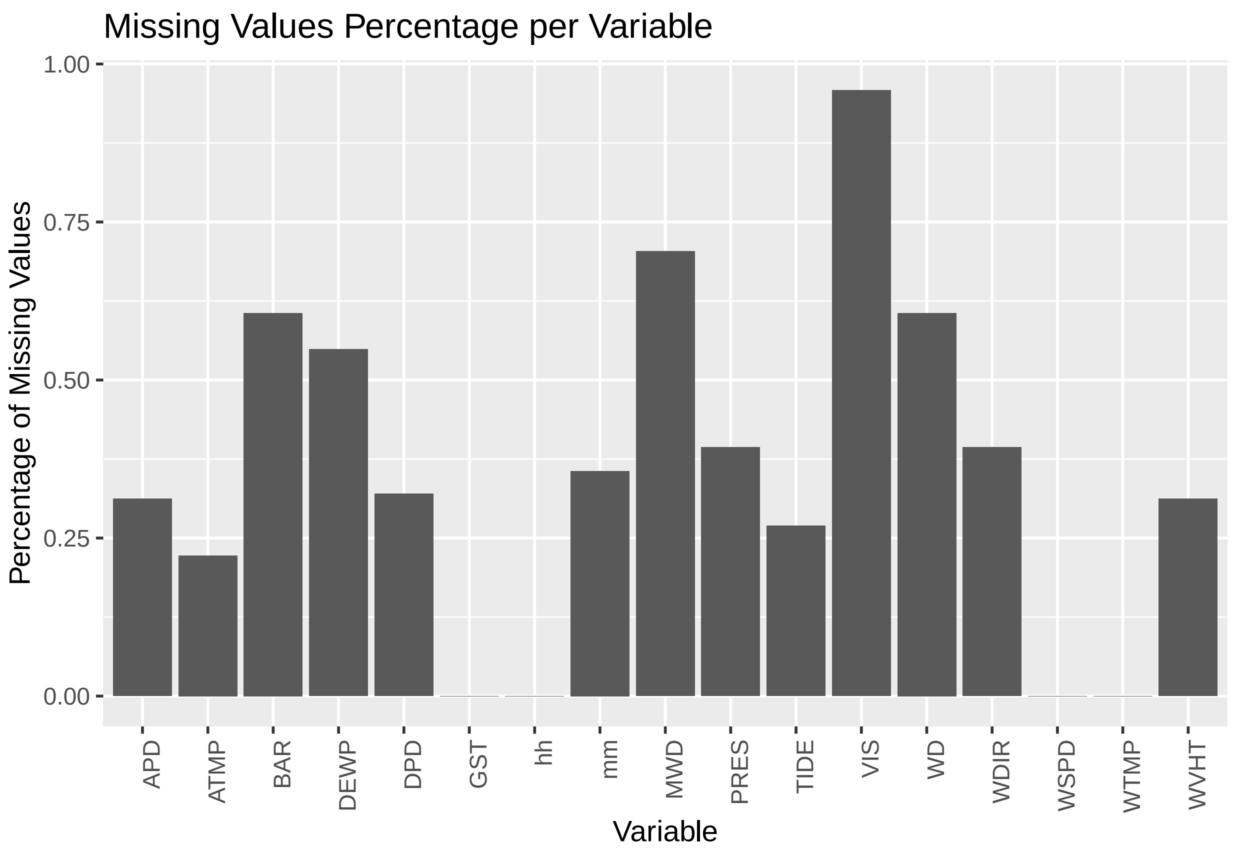
```r
all_buoy_data_2$ATMP<-ifelse(all_buoy_data_2$ATMP == 999,NA,
                             all_buoy_data_2$ATMP)
all_buoy_data_2$MWD<-ifelse(all_buoy_data_2$MWD == 999,NA,
                            all_buoy_data_2$MWD)
all_buoy_data_2$APD<-ifelse(all_buoy_data_2$APD == 99,NA,
                            all_buoy_data_2$APD)
all_buoy_data_2$DPD<-ifelse(all_buoy_data_2$DPD == 99,NA,
                            all_buoy_data_2$DPD)
all_buoy_data_2$WVHT<-ifelse(all_buoy_data_2$WVHT == 99,NA,
                             all_buoy_data_2$WVHT)
all_buoy_data_2$DEWP<-ifelse(all_buoy_data_2$DEWP == 999,NA,
                             all_buoy_data_2$DEWP)
all_buoy_data_2$VIS<-ifelse(all_buoy_data_2$VIS == 99,NA,
                            all_buoy_data_2$VIS)

buoy_clean <- all_buoy_data_2
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
missing_summary <- buoy_clean %>%
  select(-c(YY, `#YY`,DD, MM, YYYY, new_column,Year)) %>%
  summarise(across(everything(), ~mean(is.na(.), na.rm = TRUE)))
print(missing_summary)
```

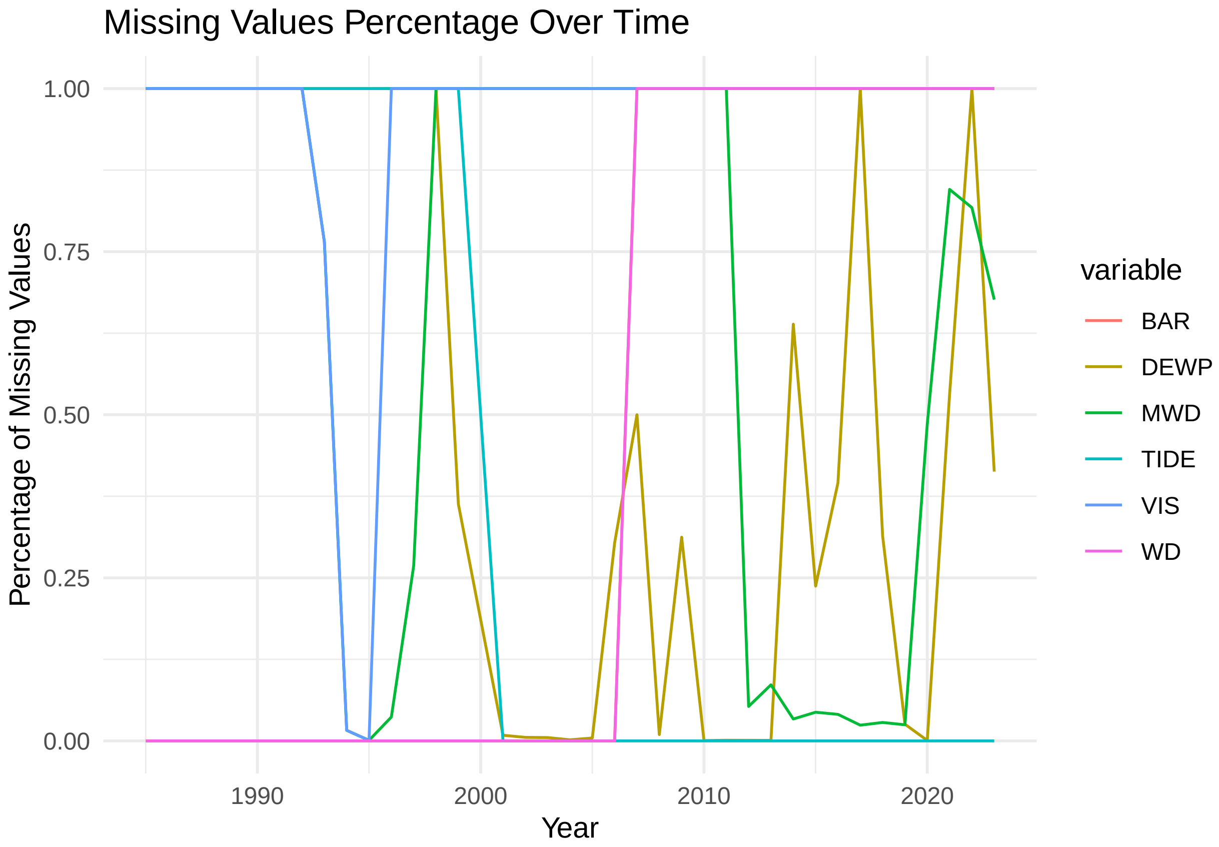
```
##   hh        WD WSPD GST      WVHT       DPD       APD       MWD       BAR
## 1  0 0.6061419    0   0 0.3120672 0.3200534 0.3120672 0.7036476 0.6061419
##        ATMP WTMP      DEWP       VIS      TIDE        mm      WDIR      PRES
## 1 0.2222816    0 0.5485885 0.9583843 0.2693007 0.3561532 0.3938581 0.3938581
```

```r
library(tidyr)
missing_long <- gather(missing_summary, key = "variable",
                       value = "missing_percentage")
library(ggplot2)
ggplot(missing_long, aes(x = variable, y = missing_percentage)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Missing Values Percentage per Variable",
       y = "Percentage of Missing Values", x = "Variable")
```



```
#Not all "999" and "99" need to be treated as missing values
#there are many cases where the true value is equal to 999 or 99
#As can be seen from the bar chart, VIS has the most missing values,
#and most variables contain missing values

columns_of_interest <- buoy_clean %>%
  select(VIS,TIDE,MWD,BAR, DEWP, WD,Year)

missing_summary_by_year <- columns_of_interest %>%
  group_by(Year) %>%
  summarise(across(everything(), ~mean(is.na(.), na.rm = TRUE)))
missing_long_by_year <- missing_summary_by_year %>%
  pivot_longer(cols = -Year, names_to = "variable", values_to = "missing_percentage")
ggplot(missing_long_by_year, aes(x = Year, y = missing_percentage, color = variable)) +
  geom_line() +
  labs(title = "Missing Values Percentage Over Time",
       y = "Percentage of Missing Values",
       x = "Year") +
  theme_minimal()
```

```
## Warning: Removed 6 rows containing missing values or values outside the scale range
## (`geom_line()`).
```



```
#According to the BAR chart, VIS,TIDE,MWD,BAR, DEWP and WD
#which had a large number of overall missing values
#were selected to plot the yearly change of the proportion of missing values
#and no obvious periodicity or trend was found
```

(c)

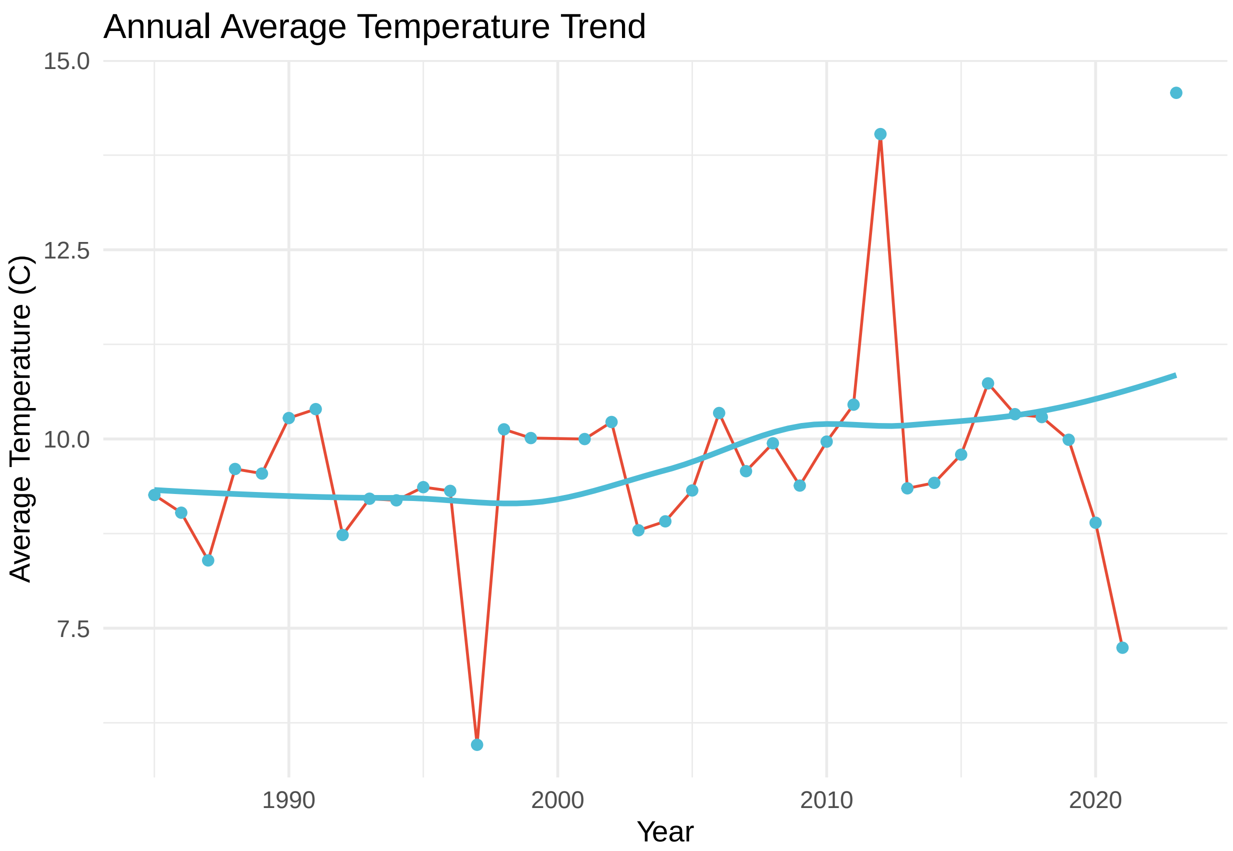
```
annual_avg_temp <- buoy_clean %>%
  group_by(Year) %>%
  summarise(avg_temp = mean(ATMP, na.rm = TRUE))

ggplot(annual_avg_temp, aes(x = Year, y = avg_temp)) +
  geom_line(color = "#E64B35") +
  geom_point(color = "#4DBBD5") +
  geom_smooth(method = "loess", color = "#4DBBD5", se = FALSE) +
  labs(title = "Annual Average Temperature Trend",
       x = "Year",
       y = "Average Temperature (C)") +
  theme_minimal()
```

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

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

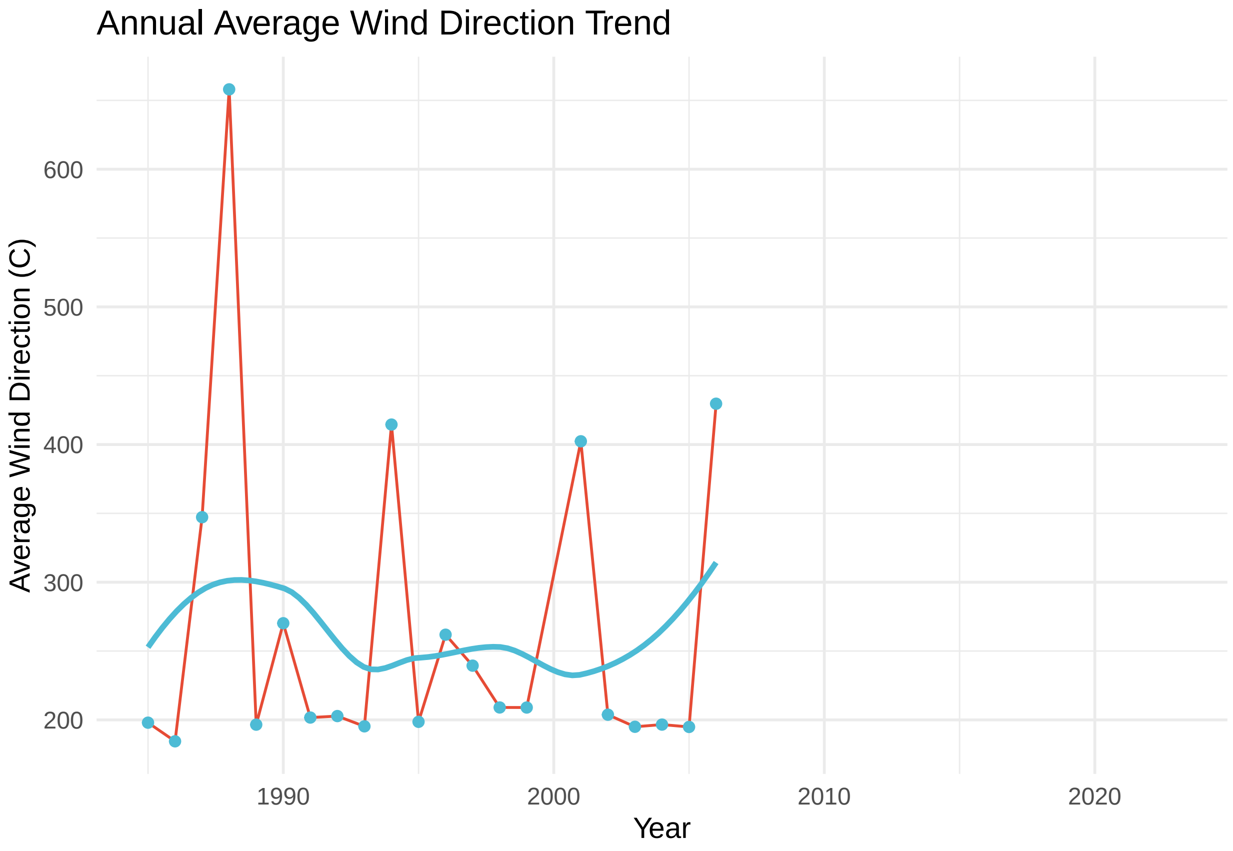


```
#According to the linear graph of average temperature changes
#and the fitted curve for each year, it can be seen that the temperature
#has shown a slow fluctuating upward trend from 1985 to 2023

annual_avg_wd <- buoy_clean %>%
  group_by(Year) %>%
  summarise(avg_wd = mean(WD, na.rm = TRUE))

ggplot(annual_avg_wd, aes(x = Year, y = avg_wd)) +
  geom_line(color = "#E64B35") +
  geom_point(color = "#4DBBD5") +
  geom_smooth(method = "loess", color = "#4DBBD5", se = FALSE) +
  labs(title = "Annual Average Wind Direction Trend",
       x = "Year",
       y = "Average Wind Direction (C)") +
  theme_minimal()
```

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

```
## Warning: Removed 18 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 18 rows containing missing values or values outside the scale range
## (`geom_line()`).
## Warning: Removed 18 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
#From the graph, it can be observed that the wind direction angles
#fluctuated significantly between 1985 and 2023, but there is no apparent trend
```

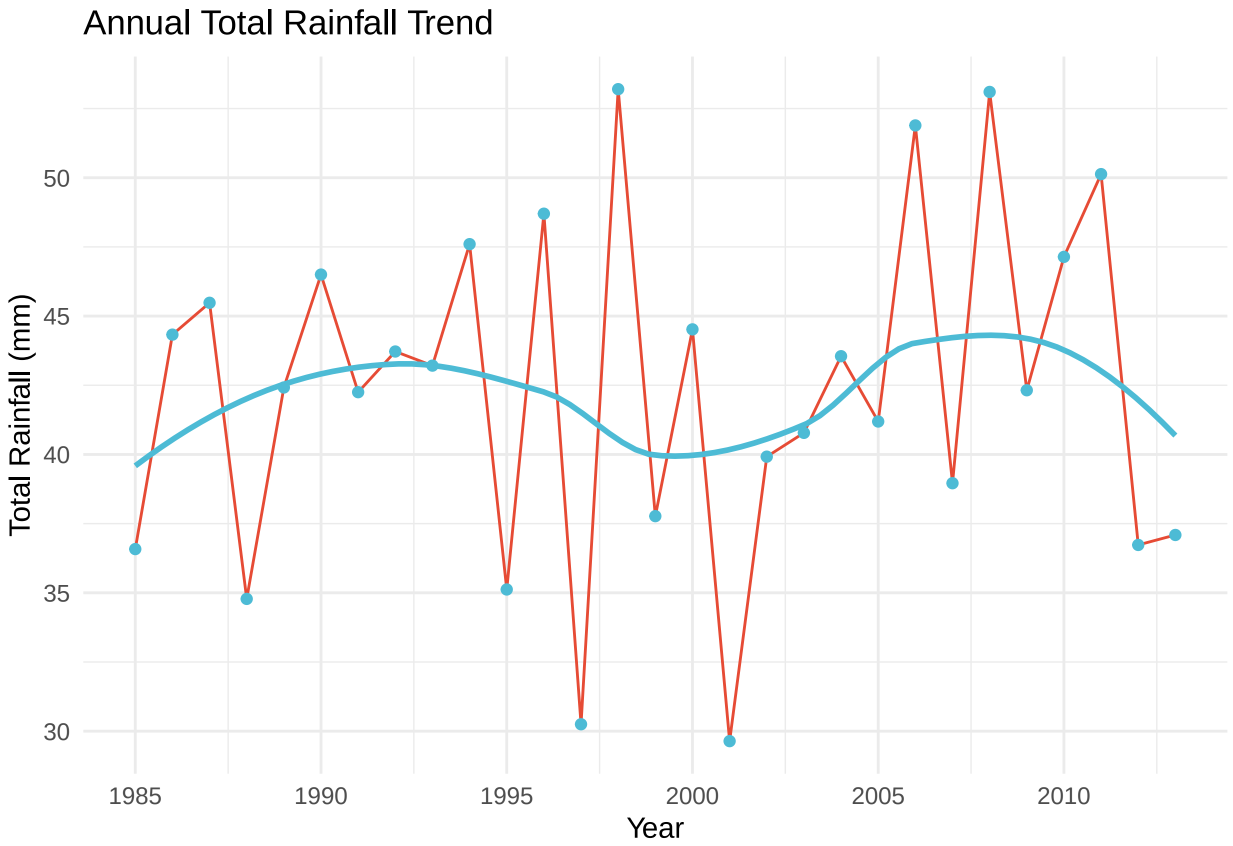
(d)

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

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

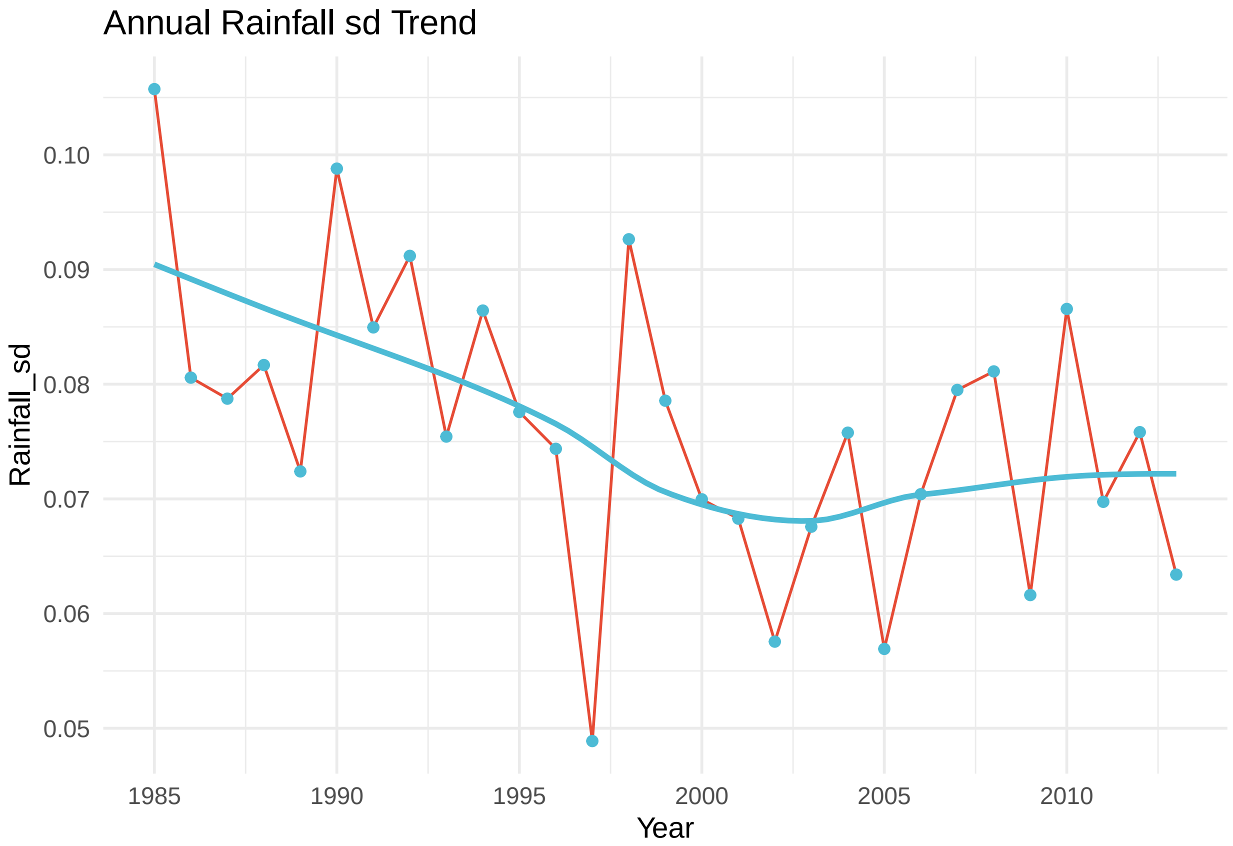
```
library(dplyr)

rain_data <- read.csv("Rainfall.csv")
rain_data$Date <- as.Date(rain_data$DATE, format = "%Y%m%d %H:%M")
rain_data <- rain_data %>%
  mutate(Year = year(Date))
annual_rainfall <- rain_data %>%
  group_by(Year) %>%
  summarise(total_rainfall = sum(HPCP, na.rm = TRUE),
            rainfall_sd = sd(HPCP, na.rm = TRUE))
ggplot(annual_rainfall, aes(x = Year, y = total_rainfall)) +
  geom_line(color = "#E64B35") +
  geom_point(color = "#4DBBD5") +
  geom_smooth(method = "loess", color = "#4DBBD5", se = FALSE) +
  labs(title = "Annual Total Rainfall Trend",
       x = "Year",
       y = "Total Rainfall (mm)") +
  theme_minimal()
```

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



```
ggplot(annual_rainfall, aes(x = Year, y = rainfall_sd)) +
  geom_line(color = "#E64B35") +
  geom_point(color = "#4DBBD5") +
  geom_smooth(method = "loess", color = "#4DBBD5", se = FALSE) +
  labs(title = "Annual Rainfall sd Trend",
       x = "Year",
       y = "Rainfall_sd") +
  theme_minimal()
```

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



```
#As can be seen from the above two plots, although the rainfall in Boston
#did not show an obvious trend from 1985 to 2013
#the standard deviation of precipitation in each year showed a decreasing trend,
#indicating that the precipitation in Boston from 1985 to 2013 had a tendency to be
#evenly distributed throughout the year
print(annual_avg_temp)
```

```
## # A tibble: 39 x 2
##     Year avg_temp
##    <int>    <dbl>
##  1  1985     9.26
##  2  1986     9.03
##  3  1987     8.40
##  4  1988     9.60
##  5  1989     9.54
##  6  1990    10.3
##  7  1991    10.4
##  8  1992     8.73
##  9  1993     9.21
## 10  1994     9.19
## # i 29 more rows
```

```
print(annual_rainfall)
```

```
## # A tibble: 29 x 3
##     Year total_rainfall rainfall_sd
##    <dbl>          <dbl>       <dbl>
##  1  1985           36.6      0.106
##  2  1986           44.3      0.0806
##  3  1987           45.5      0.0787
##  4  1988           34.8      0.0817
##  5  1989           42.4      0.0724
##  6  1990           46.5      0.0988
##  7  1991           42.2      0.0850
##  8  1992           43.7      0.0912
##  9  1993           43.2      0.0754
## 10  1994           47.6      0.0864
## # i 19 more rows
```

```
annual_avg_temp_filtered <- annual_avg_temp %>%
  filter(Year >= 1985 & Year <= 2013)

annual_rainfall_filtered <- annual_rainfall %>%
  filter(Year >= 1985 & Year <= 2013)
combined_data <- inner_join(annual_avg_temp_filtered, annual_rainfall_filtered, by = "Year")
model <- lm(avg_temp ~ total_rainfall, data = combined_data)
summary(model)
```

```
##
## Call:
## lm(formula = avg_temp ~ total_rainfall, data = combined_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2545 -0.5076 -0.0309  0.4686  4.6085
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.25457    1.62282   5.087 2.67e-05 ***
## total_rainfall  0.03172    0.03797   0.836    0.411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 26 degrees of freedom
## Multiple R-squared:  0.02615,    Adjusted R-squared:  -0.01131
## F-statistic: 0.6981 on 1 and 26 DF,  p-value: 0.411
```

```
#The model performs poorly,
#with precipitation showing no statistically significant impact on annual average temperature,
#and the explanatory power is very low (R-squared is only 2.6%).
#This model is nearly ineffective for explaining the variation in annual average temperature.
```