HW 4

Zilu Sun

Fall 2024

(a)

##Your first exercise is to read in the data for all the years from 1985 to 2023. As discussedin class, you don't want to do this manually and will need to figure out a way to do itprogrammatically. We've given you a skeleton of how to do this for data for one year below.Your task is to adapt this to reading in multiple datasets from all the years in question. Thisexample code is meant to be a guide and if you think of a better way to read the data in, go for it

```
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)

##Your next exercise is to identify and deal with the null data in the dataset.Recall from classthat for WDIR and some other variables these showed up as 999 in the dataset. Convert themto NA's. Is it always appropriate to convert missing/null data to NA's? When might it notbe? Analyze the pattern of NA's. Do you spot any patterns in the way/dates that these are distributed?

```
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
```

```
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
```

```
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)
```

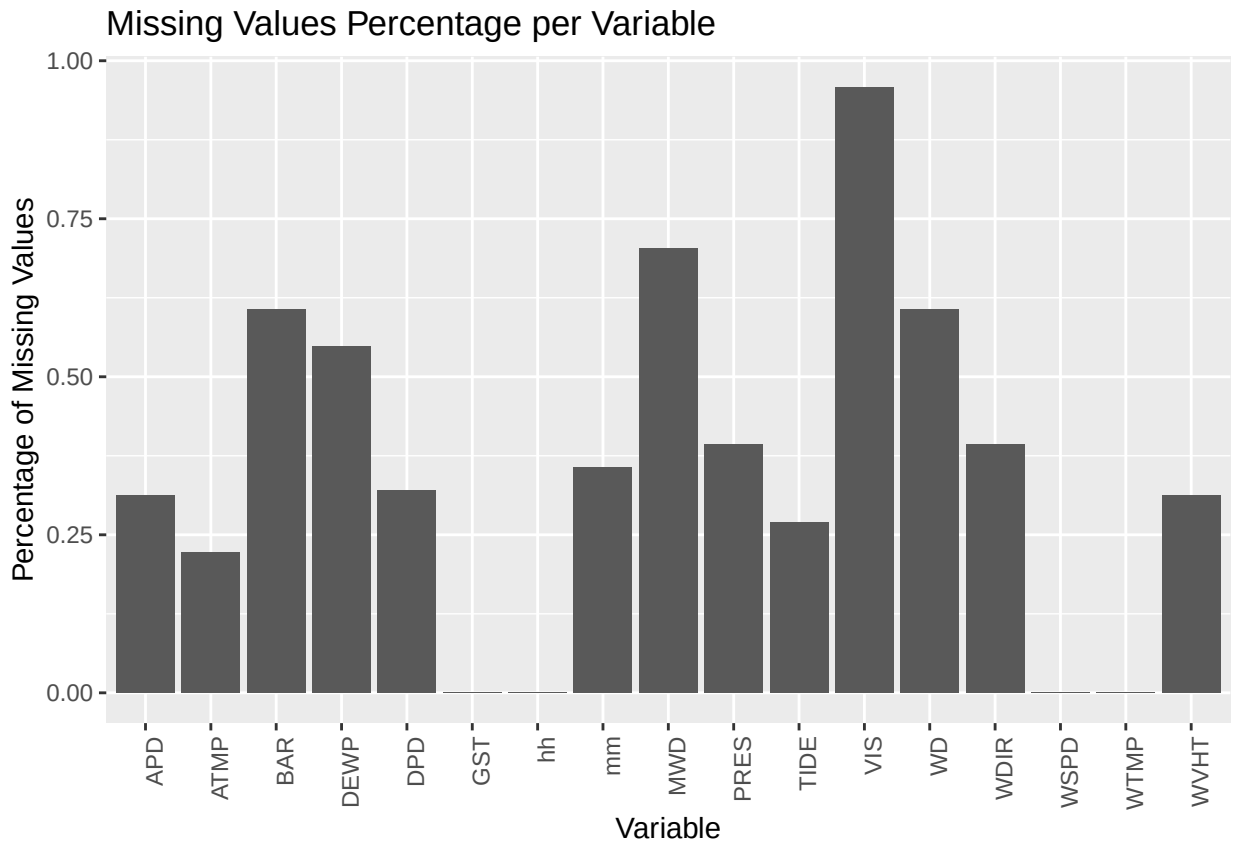
```
##
## Attaching package: 'dplyr'

## 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
```

```
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
```

```
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 = "Variab
```



(c)

Can you use the Buoy data to see the effects of climate change? Create visualizations to show this and justify your choices. Can you think of statistics you can use to bolster what your plots represent? Calculate these, justify your use of them. Add this code, its output,your answers and visualizations to your pdf.

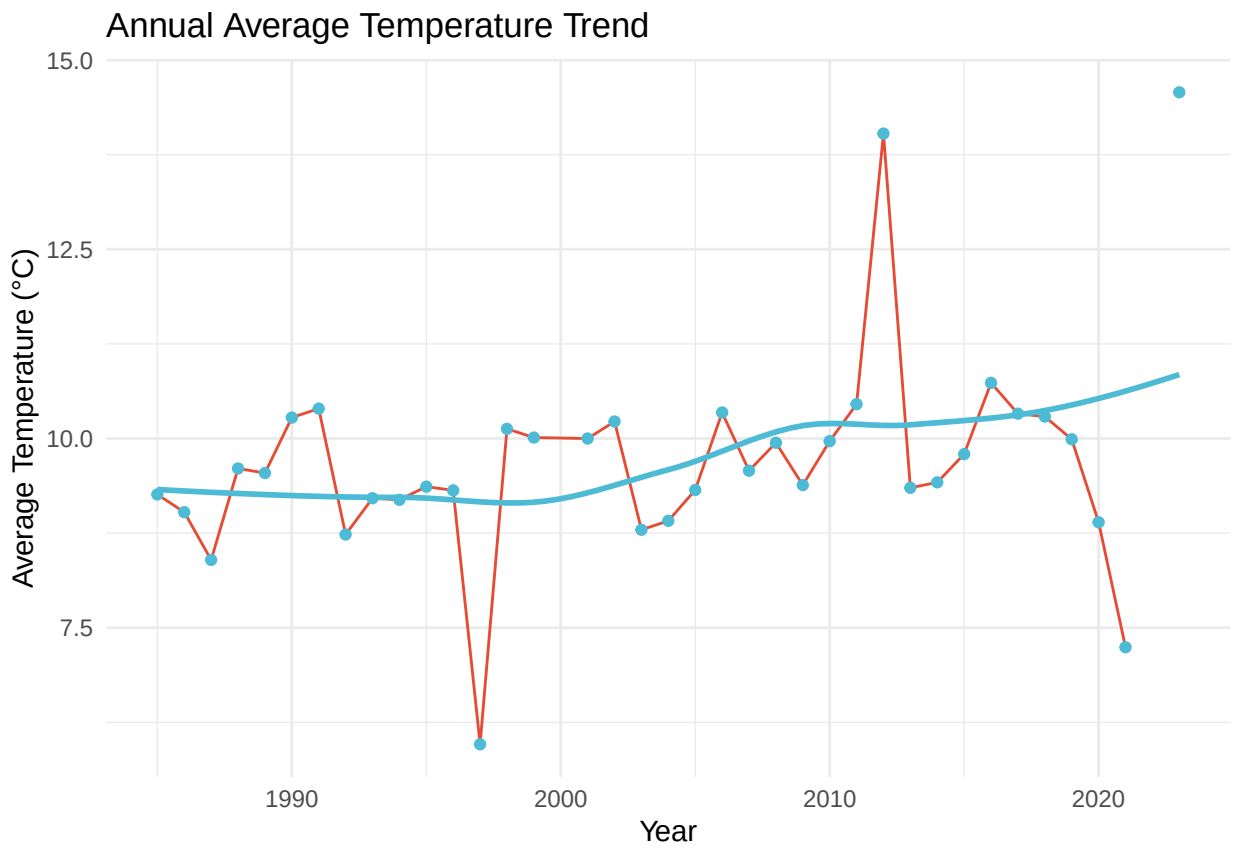
```
annual_avg_temp <- buoy_clean %>%
  group_by(Year) %>%
  summarise(avg_temp = mean(ATMP, na.rm = TRUE))

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()`).
```

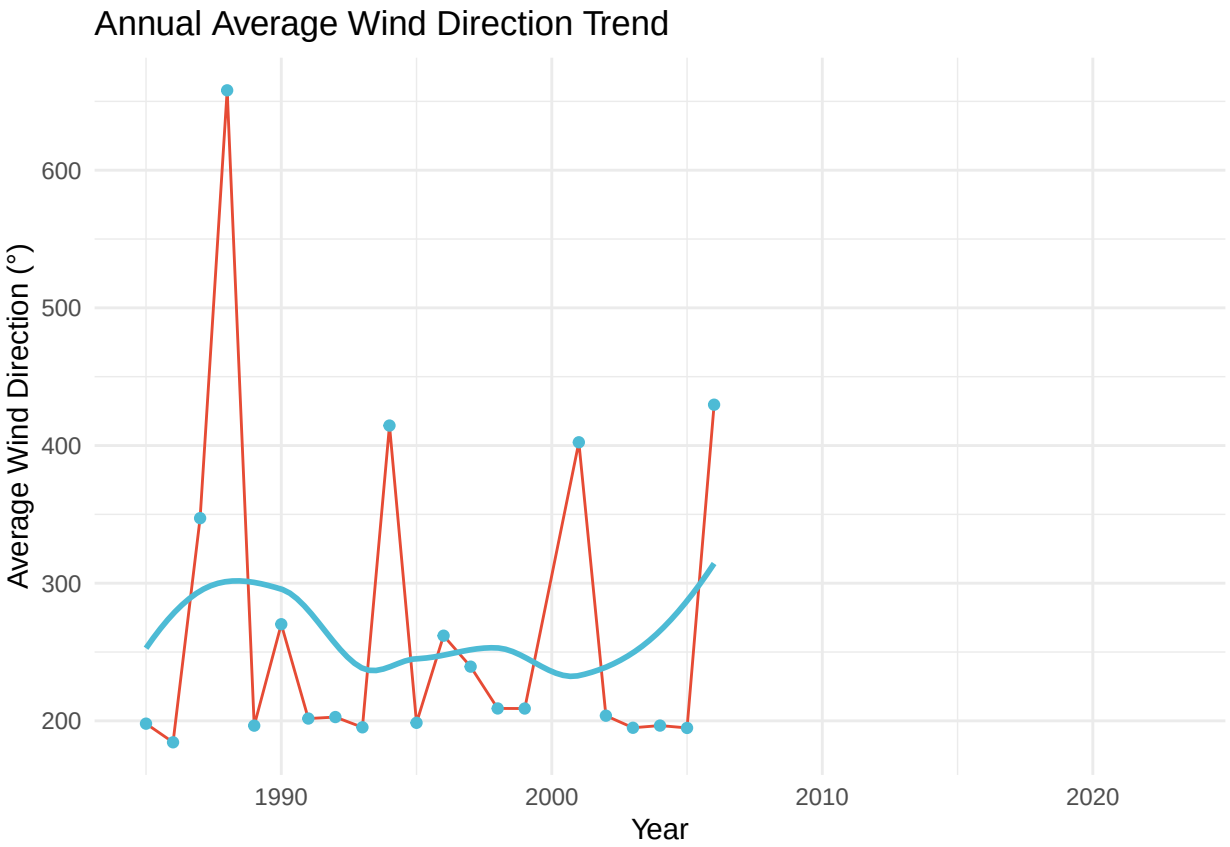
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## 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 ca

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 (°)") +
  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 198
```

(d)

```
library(lubridate)
```

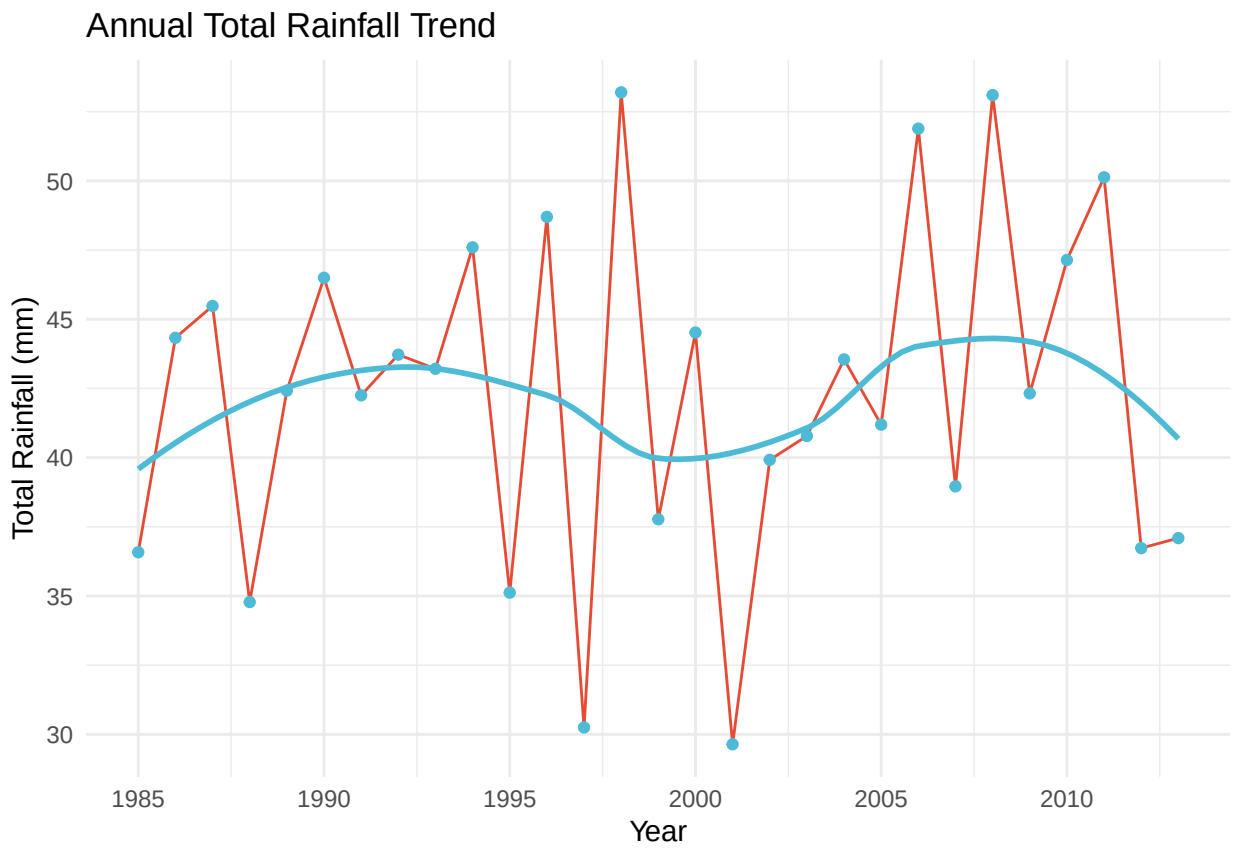
```
##
## Attaching package: 'lubridate'

## 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
```

```r
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))
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
```

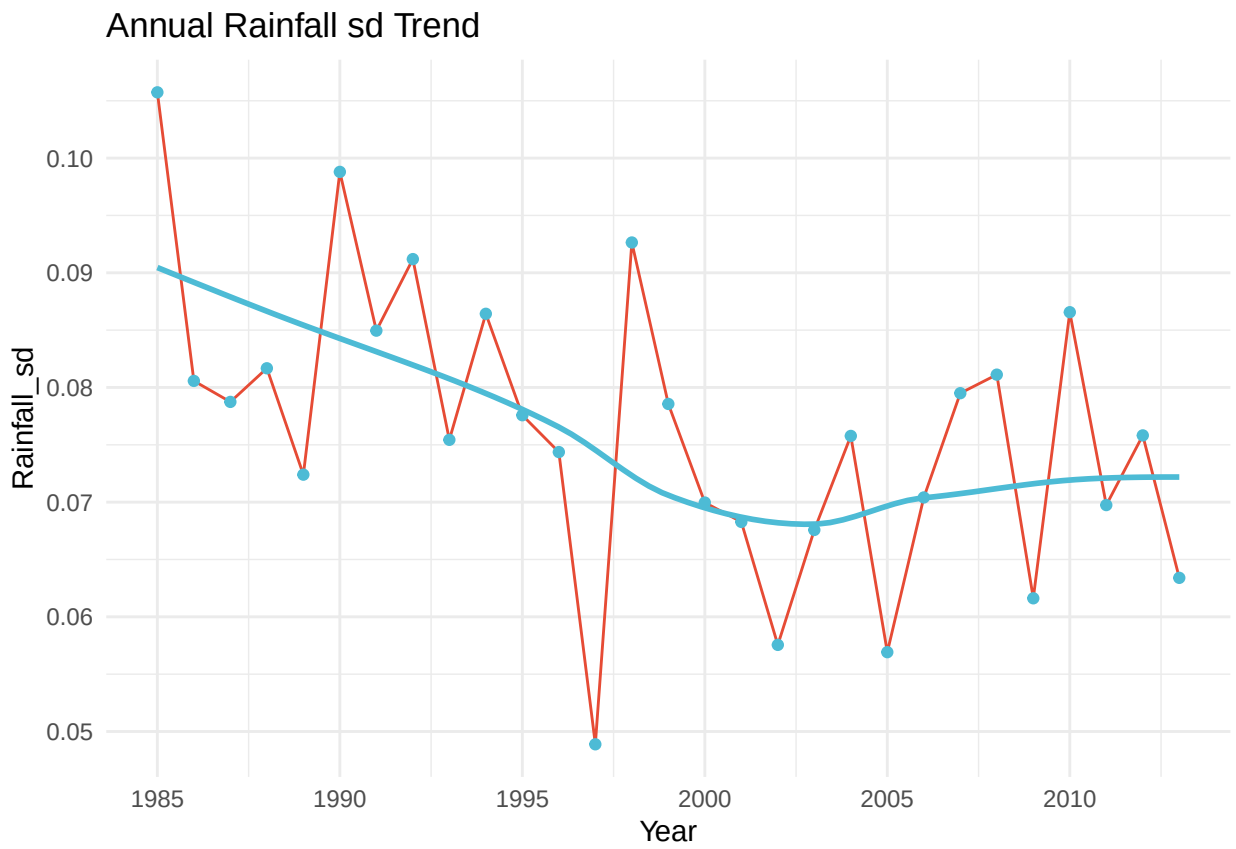
```r
ggplot(annual_rainfall, aes(x = Year, y = total_rainfall)) +
  geom_line(color = "#E64B35") +
  geom_point(color = "#4DBBD5") +
  geom_smooth(method = "loess", color = "#4DBBD5", se = FALSE) +
  labs(title = "Annual Total Rainfall Trend",
       x = "Year",
       y = "Total Rainfall (mm)") +
  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
```