Buoy_Students

Packages to Install

```
library(data.table)
library(dplyr)
library(lubridate)
library(ggplot2)
library(zoo)
library(tibble)
library(readr)
```

Compiling all years' data

1. Try to remember why 2007 is the split year
2. Why are using different functions to read data prior and post 2007?
3. What are some new function you got to know from following code? what do these functions do?

```
file_root <- "https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
tail <- ".txt.gz&dir=data/historical/stdmet/"

load_buoy_data <- function(year) {
  path <- paste0(file_root, year, tail)

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

  if (num_columns == 16) {
    buoy <- read.table(path, fill = TRUE, header = TRUE, sep = "")
    buoy <- add_column(buoy, mm = NA, .after = "hh")
    buoy <- add_column(buoy, TIDE = NA, .after = "VIS")
```

```
  } else if (num_columns == 17) {
    buoy <- read.table(path, fill = TRUE, header = TRUE, sep = "")
    buoy <- add_column(buoy, TIDE = NA, .after = "VIS")

  } else {
    buoy <- fread(path, header = FALSE, skip = 1, fill = TRUE)
    setnames(buoy, header)
  }

  return(buoy)
}
all_data <- lapply(1985:2024, load_buoy_data)
combined_data <- rbindlist(all_data, fill = TRUE)
```

Cleaning and Organizing the data

We start by merging all different version of `Year` column. We do same with other columns which are same but having data for certain set of years. We remove the remaining columns after merging them.

Creating datetime column using `lubridate()`

```
combined_data <- combined_data %>%
  mutate(
    YY = as.character(YY),
    `#YY` = as.character(`#YY`),
    YYYY = as.character(YYYY)
  )

# Combine year columns safely using coalesce
combined_data <- combined_data %>%
  mutate(YYYY = coalesce(YYYY, `#YY`, YY))
combined_data <- combined_data %>%
  mutate(BAR = coalesce(as.numeric(BAR), as.numeric(PRES)),  # Convert BAR and PRES to numer
    WD = coalesce(as.numeric(WD), as.numeric(WDIR)))
```

```
Warning: There were 2 warnings in `mutate()`.
The first warning was:
i In argument: `BAR = coalesce(as.numeric(BAR), as.numeric(PRES))`.
Caused by warning in `list2()`:
! NAs introduced by coercion
```

```
i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```
combined_data <- combined_data %>%
  select(-TIDE, -TIDE.1, -mm,- WDIR, -PRES,-`#YY`,-YY)

combined_data$datetime <- ymd_h(paste(combined_data$YYYY, combined_data$MM, combined_data$DD
```

```
Warning: 18 failed to parse.
```

```
combined_data <- combined_data %>%
  mutate(across(everything(),
                ~ na_if(as.numeric(as.character(.)), 99) %>%
                na_if(999) %>%
                na_if(9999)))
```

```
Warning: There were 15 warnings in `mutate()`.
The first warning was:
i In argument: `across(...)`.
Caused by warning in `vec_cast()`:
! NAs introduced by coercion
i Run `dplyr::last_dplyr_warnings()` to see the 14 remaining warnings.
```

```
summary(combined_data)
```

```
       MM               DD               hh               WD        
 Min.   : 1.000   Min.   : 1.00   Min.   : 0.0   Min.   :  0.0  
 1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 6.0   1st Qu.:128.0  
 Median : 7.000   Median :16.00   Median :11.0   Median :206.0  
 Mean   : 6.609   Mean   :15.73   Mean   :11.5   Mean   :196.8  
 3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:17.0   3rd Qu.:279.0  
 Max.   :12.000   Max.   :31.00   Max.   :23.0   Max.   :360.0  
 NA's   :18       NA's   :18      NA's   :18     NA's   :44323  
      WSPD             GST              WVHT             DPD        
 Min.   : 0.000   Min.   : 0.000   Min.   :0.00     Min.   : 0.000  
 1st Qu.: 3.400   1st Qu.: 4.200   1st Qu.:0.41     1st Qu.: 4.550  
 Median : 5.300   Median : 6.400   Median :0.66     Median : 7.690  
 Mean   : 5.864   Mean   : 7.269   Mean   :0.87     Mean   : 7.384  
 3rd Qu.: 7.900   3rd Qu.: 9.700   3rd Qu.:1.06     3rd Qu.:10.000  
 Max.   :25.700   Max.   :32.400   Max.   :9.10     Max.   :25.000  
 NA's   :33236    NA's   :33538    NA's   :179542   NA's   :183550  
```

```
      APD              MWD              BAR             ATMP        
 Min.   : 0.000   Min.   :  0.0    Min.   : 964.6   Min.   :-19.700  
 1st Qu.: 3.830   1st Qu.: 78.0    1st Qu.:1010.5   1st Qu.:  4.000  
 Median : 4.690   Median : 94.0    Median :1015.8   Median :  9.800  
 Mean   : 4.935   Mean   :124.2    Mean   :1015.8   Mean   :  9.973  
 3rd Qu.: 5.810   3rd Qu.:129.0    3rd Qu.:1021.3   3rd Qu.: 16.800  
 Max.   :12.100   Max.   :360.0    Max.   :1047.0   Max.   : 32.100  
 NA's   :179542   NA's   :366695   NA's   :6893     NA's   :102822   
      WTMP             DEWP             VIS             YYYY     
 Min.   :-1.80   Min.   :-24.900   Min.   : 0.00    Min.   :  85  
 1st Qu.: 5.90   1st Qu.: -0.100   1st Qu.: 8.10    1st Qu.:2000  
 Median :10.60   Median :  7.100   Median : 9.40    Median :2016  
 Mean   :11.16   Mean   :  6.698   Mean   :12.48    Mean   :1585  
 3rd Qu.:16.50   3rd Qu.: 14.700   3rd Qu.:11.60    3rd Qu.:2022  
 Max.   :27.80   Max.   : 26.100   Max.   :36.00    Max.   :2024  
 NA's   :16132   NA's   :253681    NA's   :499417   NA's   :18    
    datetime    
 Min.   : NA    
 1st Qu.: NA    
 Median : NA    
 Mean   :NaN    
 3rd Qu.: NA    
 Max.   : NA    
 NA's   :518656 
```

```r
#str(combined_data)
#str(combined_data$datetime)
if (!inherits(combined_data$datetime, "POSIXct")) {
  combined_data$datetime <- ymd_h(paste(combined_data$YYYY, combined_data$MM, combined_data$
}
```

```
Warning: 18 failed to parse.
```

```r
combined_data <- combined_data %>%
  mutate(Year = year(datetime))

combined_data <- combined_data %>% select(-YYYY)

yearly_avg_temp <- combined_data %>%
  group_by(Year) %>%
  summarise(
    avg_air_temp = mean(ATMP, na.rm = TRUE),
```

```
    avg_water_temp = mean(WTMP, na.rm = TRUE)
  )

# Ensure Year is numeric
yearly_avg_temp$Year <- as.numeric(as.character(yearly_avg_temp$Year))

# ggplot(yearly_avg_temp, aes(x = Year)) +
#   geom_line(aes(y = avg_air_temp, color = "Air Temperature"), size = 1) +
#   geom_line(aes(y = avg_water_temp, color = "Water Temperature"), size = 1) +
#   scale_x_continuous(
#     breaks = seq(min(yearly_avg_temp$Year, na.rm = TRUE),
#                  max(yearly_avg_temp$Year, na.rm = TRUE),
#                  2)   # every year
#   ) +
#   labs(
#     title = "Temperature Trends Over Time",
#     x = "Year",
#     y = "Temperature (°C)",
#     color = "Legend"
#   ) +
#   theme_minimal() +
#   theme(
#     axis.text.x = element_text(angle = 45, hjust = 1)
#   )
```

Question: Is there a correlation between WSPD and WVHT?

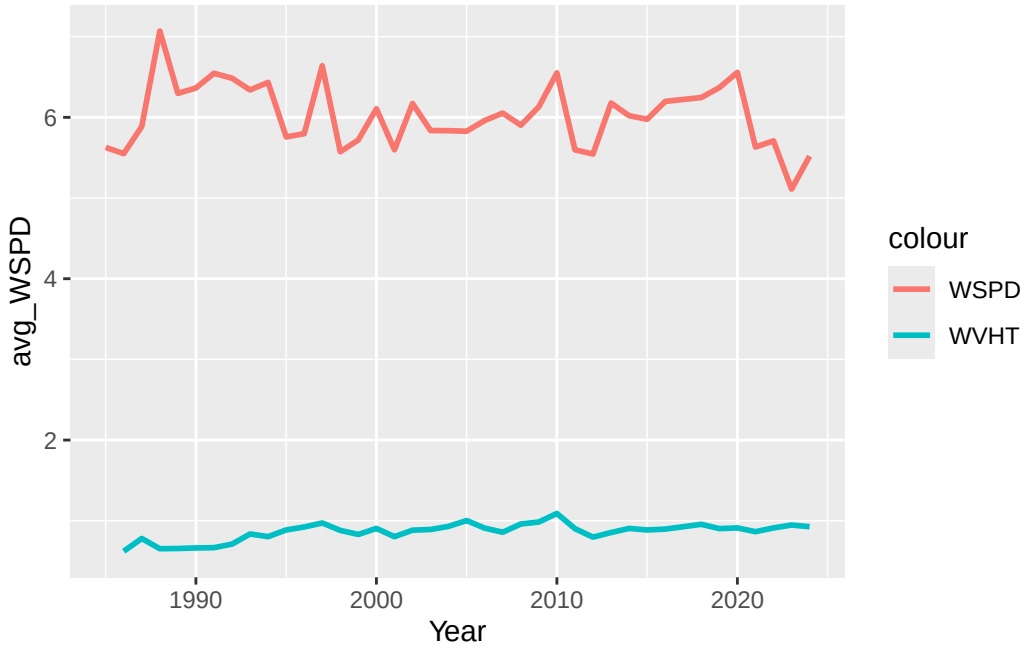
```
yearly_avg_data <- combined_data %>%
  group_by(Year) %>%
  summarise(
    avg_WSPD = mean(WSPD, na.rm = TRUE),
    avg_WVHT = mean(WVHT, na.rm = TRUE),
    avg_BAR = mean(BAR,na.rm = TRUE)
    )

ggplot(data = yearly_avg_data,aes(x = Year))+
  geom_line(aes(y = avg_WSPD,color = "WSPD"),size = 1) +
  geom_line(aes(y = avg_WVHT,color = "WVHT"),size = 1)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
```

```
i Please use `linewidth` instead.
```

```
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_line()`).
```



#We can see a roughly positive relationship between these two elements.

#We need to use scatter point and linear model to find out the result based on 2024 data

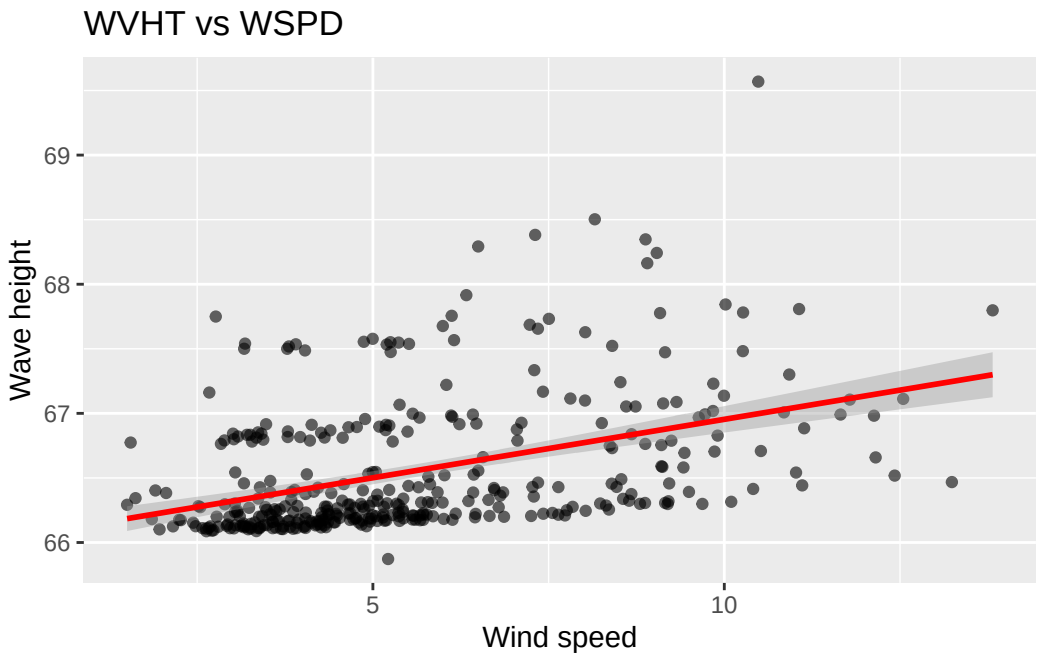
```
df2024 <- read.csv("2024.txt",sep = "",header = TRUE)
df2024 <- df2024[-1,]
df2024 <- df2024 %>%
  mutate(
    date_label = paste0(as.integer(MM), ".", as.integer(DD)),
    WSPD = as.numeric(WSPD),
    WVHT = as.numeric(WVHT)
  )
#clean up the data set and make wspd and wvht numeric

df2024_daily <- df2024 %>%
```

```r
  group_by(date_label) %>%
  summarise(
    avg_WSPD = mean(WSPD, na.rm = TRUE),
    avg_WVHT = mean(WVHT, na.rm = TRUE),
    )

fit <- lm(avg_WVHT ~ avg_WSPD,data = df2024_daily)
summary(fit)
```

```
Call:
lm(formula = avg_WVHT ~ avg_WSPD, data = df2024_daily)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7790 -0.2905 -0.1782  0.1940  2.5716

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.04891    0.06250 1056.73   <2e-16 ***
avg_WSPD     0.09048    0.01029    8.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4792 on 364 degrees of freedom
Multiple R-squared:  0.1751,    Adjusted R-squared:  0.1728
F-statistic: 77.27 on 1 and 364 DF,  p-value: < 2.2e-16
```

```r
ggplot(df2024_daily, aes(x = avg_WSPD, y = avg_WVHT)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "WVHT vs WSPD",x = "Wind speed",y = "Wave height")
```

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



From the plot, we can tell there's a positive relationship between WVHT and WSPD