---
title: "HW4 b-d"
NAME: Yishun Zhang
output: pdf_document
date: "2024-09-22"

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(data.table)
library(ggplot2)
library(dplyr)
library(lubridate)

#a. read data for years from 1985 to 2023.
setwd("C:/Users/16597/Downloads")
#
rainfall<- read.csv("Rainfall.csv", header = TRUE)
file_root <- "https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
tail <- ".txt.gz&dir=data/historical/stdmet/"

final <- data.table()

for (year in 1985:2023) {
  path <- paste0(file_root, year, tail)
  header <- scan(path, what = 'character', nlines = 1, quiet = TRUE)
  skip_lines <- ifelse(year >= 2007, 2, 1)
  buoy_data <- fread(path, header = FALSE, skip = skip_lines, fill=Inf)
  actual_col_count <- ncol(buoy_data)
  header_col_count <- length(header)

  if (header_col_count > actual_col_count) {
    header <- header[1:actual_col_count]
  } else if (header_col_count < actual_col_count) {
    header <- c(header, paste0("V", (header_col_count + 1):actual_col_count))
  }

  setnames(buoy_data, header)
  buoy_data$Year <- year
  buoy_data$Date <- make_datetime(
    year = buoy_data$YY + ifelse(buoy_data$YY >= 50, 1900, 2000),
    month = buoy_data$MM,
    day = buoy_data$DD,
    hour = buoy_data$hh,
  )
```

```
  final <- rbind(final, buoy_data, fill = TRUE)
}
```

Problem b
```{r}
missvalue <- c("WDIR", "MWD", "DEWP")

final[, (missvalue) := lapply(.SD, function(x) replace(x, x == 999, NA)), .SDcols = missvalue]
final

NAsummary <- final[, lapply(.SD, function(x) sum(is.na(x))), by = Year, .SDcols = missvalue]

print(NAsummary)

ggplot(data = NAsummary, aes(x = as.numeric(Year)))+
    geom_point(aes(y = DEWP, color = "DEWP"), na.rm = TRUE)+
    geom_line(aes(y = DEWP, color = "DEWP", group = 1), na.rm = TRUE)+
    geom_point(aes(y = MWD, color = "MWD"), na.rm = TRUE)+
    geom_line(aes(y = MWD, color = "MWD", group = 1), na.rm = TRUE) +
    geom_point(aes(y = WDIR, color = "WDIR"), na.rm = TRUE)+
    geom_line(aes(y = WDIR, color = "WDIR", group = 1), na.rm = TRUE)+
    labs(x = "Year", y = "Missing Values")+
    theme_minimal()
```
Discussion: It is always appropriate to convert missing/null data to NA's, but if 999 is a data measurement, we can't convert it.

#Problem c

```{r}
watertemp<- final[, .(meanWTMP = mean(WTMP, na.rm = TRUE)), by = Year]
watertemp

final$WTMP<- ifelse(final$WTMP ==999, NA, final$WTMP)

ggplot(data= watertemp, aes(x= Year, y= meanWTMP))+geom_point()+geom_line()+theme_minimal()

temptrend<- lm(Year~meanWTMP, data = watertemp)
temptrend

ggplot(data = watertemp, aes(x = Year, y = meanWTMP)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue")+
```

```
  labs(x = "Year", y = "Mean WTMP")+
  theme_minimal()
```
#Problem d
step1: Create summaries for rainfall and buoy data.
```{r}
str(final)
str(rainfall)

rainsummary<- rainfall %>%
  summarise(
  meanrain= mean(rainfall$HPCP, na.rm= TRUE),
  medianrain= median(rainfall$HPCP, na.rm= TRUE),
  countrain= sum(!is.na(rainfall$HPCP)))

rainsummary

buoysummary<- final%>%
  summarise(meantemp= mean(WTMP, na.rm = TRUE),
            mediantemp= median(WTMP, na.rm= TRUE),
            meanwind= mean(WSPD, na.rm= TRUE),
            medianwind= median(WSPD, na.rm= TRUE))

buoysummary
```


Step2: Find relationships among HPCP, WTMP and WSPD. Make visualizations for them.
```{r}
#Set the same form of the date
rainfall$Date<- as.Date(rainfall$DATE, format= "%Y%m%d %H:%M")
final$Date<- as.Date(final$Date, format= "%Y%m%d %H:%M")

combinedata<- merge(rainfall, final, by= "Date")
combinedata

combinedata <- combinedata %>%
  filter(!is.na(HPCP), !is.na(WTMP), !is.na(WSPD))

#Create the visualization for WTMP and HPCP
ggplot(combinedata, aes(x = WTMP, y = HPCP, color= "Date")) +
  geom_point(alpha = 0.5, color = "blue") +
  theme_minimal()
#Create the visualization for WSPD and HPCP
ggplot(combinedata, aes(x = WSPD, y = HPCP)) +
  geom_point(alpha = 0.5, color = "red") +
  theme_minimal()
#Create the visualization for the date and HPCP
```

```
ggplot(combinedata, aes(x = Date, y = HPCP)) +
  geom_point(color = "yellow") +
  geom_smooth(color = "blue") +
  theme_minimal()
```

Step3: Create a small linear model.

```{r}
HPCPmodel<- lm(HPCP~WTMP+WSPD, data= combinedata)# Find the linear relationship HPCP~water temperature+ wind speed with no interaction.
ggplot(combinedata, aes(x= WTMP, y= HPCP))+# Create the ggplot for the model.
geom_point(alpha = 0.5, color = "blue")+# add points to the graph
  geom_smooth(method = "lm", se = FALSE, color = "black")+ #add lines to the graph
  theme_minimal()
```