STA141B HW1 Part1

Bohan Chen

2025-04-08

Set Up File Paths and Enumerate the Zip Files

```
#Set Up File Paths and Enumerate the Zip Files
dir_path <- "~/Desktop/Solar1"
zip_files <- list.files(dir_path, pattern = "\\.zip$", full.names = TRUE)

zip_files
```

```
## [1] "/Users/chenbohan/Desktop/Solar1/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2
023.zip"
## [2] "/Users/chenbohan/Desktop/Solar1/USA_CA_Mount.Shasta.725957_TMYx.2009-2023.zi
p"
## [3] "/Users/chenbohan/Desktop/Solar1/USA_CA_Point.Arguello.994210_TMYx.2009-2023.z
ip"
## [4] "/Users/chenbohan/Desktop/Solar1/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.200
9-2023.zip"
## [5] "/Users/chenbohan/Desktop/Solar1/USA_CA_UC-Davis-University.AP.720576_TMYx.200
9-2023.zip"
```

Verify whether each unzipped folder contains a .clm file

```
#Verify whether each unzipped folder contains a .clm file
solar1_dir <- "~/Desktop/Solar1"

clm_folders <- list.dirs(solar1_dir, recursive = FALSE)

for (folder in clm_folders) {
  clm_files <- list.files(folder, pattern = "\\.clm$", full.names = TRUE)

  if (length(clm_files) == 0) {
    cat("Folder:", folder, "--> NO .clm FILE FOUND!\n")
  } else {
    cat("Folder:", folder, "--> Found .clm file(s):\n")
    cat("           ", paste(basename(clm_files), collapse = ", "), "\n")
  }
}
```

```
## Folder: /Users/chenbohan/Desktop/Solar1/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.200
9-2023 2 --> Found .clm file(s):
##             USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2023.clm
## Folder: /Users/chenbohan/Desktop/Solar1/USA_CA_Mount.Shasta.725957_TMYx.2009-2023
--> Found .clm file(s):
##             USA_CA_Mount.Shasta.725957_TMYx.2009-2023.clm
## Folder: /Users/chenbohan/Desktop/Solar1/USA_CA_Point.Arguello.994210_TMYx.2009-202
3 --> Found .clm file(s):
##             USA_CA_Point.Arguello.994210_TMYx.2009-2023.clm
## Folder: /Users/chenbohan/Desktop/Solar1/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.
2009-2023 --> Found .clm file(s):
##             USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.2009-2023.clm
## Folder: /Users/chenbohan/Desktop/Solar1/USA_CA_UC-Davis-University.AP.720576_TMYx.
2009-2023 --> Found .clm file(s):
##             USA_CA_UC-Davis-University.AP.720576_TMYx.2009-2023.clm
```

Define the Function

```
#Define the Function
readClmFile <- function(file_path) {
  lines <- readLines(file_path)

  n <- length(lines)
  i <- 1
  results_list <- list()

  while (i <= n) {
    line <- lines[i]

    if (grepl("^\\* day\\s+\\d+ month\\s+\\d+", line)) {
      match_vec <- regmatches(line, regexec("^\\* day\\s+(\\d+) month\\s+(\\d+)", lin
e))[[1]]
      day_val   <- as.integer(match_vec[2])
      month_val <- as.integer(match_vec[3])

      data_start <- i + 1
      data_end   <- i + 24
      if (data_end > n) {
        stop("Not enough lines for day=", day_val, ", month=", month_val)
      }

      data_lines <- lines[data_start:data_end]
      mat <- do.call(rbind, lapply(data_lines, function(x) {
        as.numeric(strsplit(x, ",")[[1]])
      }))

      hours <- 0:23
      df_day <- data.frame(
        day        = rep(day_val, 24),
        month      = rep(month_val, 24),
        hour       = hours,
        diffuse    = mat[,1],
        temp_tenths= mat[,2],
        direct     = mat[,3],
        wind_speed_tenths = mat[,4],
        wind_dir   = mat[,5],
        rh_percent = mat[,6],
        stringsAsFactors   = FALSE
      )

      results_list[[length(results_list) + 1]] <- df_day
      i <- data_end + 1
    } else {
      i <- i + 1
    }
  }

  final_df <- do.call(rbind, results_list)
  return(final_df)
}
```

After the function is defined, call it for each 5 locations

```
#After the function is defined, call it for each 5 locations
# 1) Mammoth
file_path_mammoth <- "/Users/chenbohan/Desktop/Solar1/USA_CA_Mammoth.Yosemite.AP.7238
94_TMYx.2009-2023 2/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2023.clm"
df_mammoth <- readClmFile(file_path_mammoth)
head(df_mammoth)
```

```
##   day month hour diffuse temp_tenths direct wind_speed_tenths wind_dir
## 1   1     1    0       0         -90      0                26       10
## 2   1     1    1       0         -90      0                21      310
## 3   1     1    2       0         -90      0                 0      180
## 4   1     1    3       0         -90      0                 0      187
## 5   1     1    4       0         -80      0                 0      192
## 6   1     1    5       0        -110      0                 0      172
##   rh_percent
## 1         77
## 2         70
## 3         70
## 4         64
## 5         78
## 6         70
```

```
summary(df_mammoth)
```

```
##       day            month             hour          diffuse
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00
##  Median :16.00   Median : 7.000   Median :11.50   Median : 10.00
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 53.45
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 86.00
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :512.00
##   temp_tenths         direct       wind_speed_tenths    wind_dir
##  Min.   :-248.00   Min.   :   0.0   Min.   :  0.00    Min.   :  0.0
##  1st Qu.:   4.00   1st Qu.:   0.0   1st Qu.:  0.00    1st Qu.:165.0
##  Median :  51.00   Median :   0.0   Median : 26.00    Median :227.0
##  Mean   :  67.21   Mean   : 317.5   Mean   : 32.05    Mean   :220.3
##  3rd Qu.: 130.00   3rd Qu.: 743.0   3rd Qu.: 46.00    3rd Qu.:280.0
##  Max.   : 350.00   Max.   :1080.0   Max.   :227.00    Max.   :360.0
##    rh_percent
##  Min.   : 10.00
##  1st Qu.: 34.00
##  Median : 51.00
##  Mean   : 51.14
##  3rd Qu.: 68.00
##  Max.   :100.00
```

```
# 2) UC Davis
file_path_ucdavis <- "/Users/chenbohan/Desktop/Solar1/USA_CA_UC-Davis-University.AP.7
20576_TMYx.2009-2023/USA_CA_UC-Davis-University.AP.720576_TMYx.2009-2023.clm"
df_ucdavis <- readClmFile(file_path_ucdavis)
head(df_ucdavis)
```

```
##   day month hour diffuse temp_tenths direct wind_speed_tenths wind_dir
## 1   1     1    0       0          50      0                31       30
## 2   1     1    1       0          58      0                15       80
## 3   1     1    2       0          40      0                 0        2
## 4   1     1    3       0          39      0                21      120
## 5   1     1    4       0          36      0                 0       29
## 6   1     1    5       0          33      0                 0       46
##   rh_percent
## 1         42
## 2         47
## 3         58
## 4         56
## 5         56
## 6         61
```

```
summary(df_ucdavis)
```

```
##       day            month             hour          diffuse      
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  9.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 47.86  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 88.25  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :432.00  
##   temp_tenths        direct      wind_speed_tenths    wind_dir    
##  Min.   :-10.0   Min.   :  0.0   Min.   :  0.00    Min.   :  0.0  
##  1st Qu.:110.0   1st Qu.:  0.0   1st Qu.:  6.00    1st Qu.:170.0  
##  Median :154.0   Median :  0.0   Median : 21.00    Median :210.0  
##  Mean   :168.3   Mean   :282.4   Mean   : 24.05    Mean   :215.9  
##  3rd Qu.:220.0   3rd Qu.:667.0   3rd Qu.: 36.00    3rd Qu.:292.0  
##  Max.   :390.0   Max.   :994.0   Max.   :129.00    Max.   :360.0  
##    rh_percent    
##  Min.   : 10.00  
##  1st Qu.: 37.00  
##  Median : 59.00  
##  Mean   : 58.01  
##  3rd Qu.: 78.00  
##  Max.   :100.00  
```

```
# 3) San Diego (Miramar)
file_path_sandiego <- "/Users/chenbohan/Desktop/Solar1/USA_CA_San.Diego-MCAS.Miramar.
722930_TMYx.2009-2023/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.2009-2023.clm"
df_sandiego <- readClmFile(file_path_sandiego)
head(df_sandiego)
```

```
##   day month hour diffuse temp_tenths direct wind_speed_tenths wind_dir
## 1   1     1    0       0         266      0                36       70
## 2   1     1    1     110         268    460                46       70
## 3   1     1    2     138         269    576                41       80
## 4   1     1    3     165         270    681                31       80
## 5   1     1    4     200         270    693                31       80
## 6   1     1    5     207         270    661                41       70
##   rh_percent
## 1         71
## 2         63
## 3         60
## 4         58
## 5         77
## 6         62
```

```
summary(df_sandiego)
```

```
##       day            month             hour          diffuse      
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  3.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 75.12  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.:151.25  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :429.00  
##   temp_tenths        direct      wind_speed_tenths    wind_dir    
##  Min.   :172.0   Min.   :  0.0   Min.   :  0.00    Min.   : 10.0  
##  1st Qu.:234.0   1st Qu.:  0.0   1st Qu.: 15.00    1st Qu.:147.0  
##  Median :258.0   Median :  0.0   Median : 26.00    Median :208.0  
##  Mean   :250.1   Mean   :194.4   Mean   : 25.24    Mean   :206.9  
##  3rd Qu.:271.0   3rd Qu.:453.2   3rd Qu.: 36.00    3rd Qu.:290.0  
##  Max.   :322.0   Max.   :891.0   Max.   :113.00    Max.   :360.0  
##    rh_percent    
##  Min.   : 54.00  
##  1st Qu.: 74.00  
##  Median : 80.00  
##  Mean   : 79.07  
##  3rd Qu.: 84.00  
##  Max.   :100.00  
```

```
# 4) Mount Shasta
file_path_shasta <- "/Users/chenbohan/Desktop/Solar1/USA_CA_Mount.Shasta.725957_TMYx.
2009-2023/USA_CA_Mount.Shasta.725957_TMYx.2009-2023.clm"
df_shasta <- readClmFile(file_path_shasta)
head(df_shasta)
```

```
##   day month hour diffuse temp_tenths direct wind_speed_tenths wind_dir
## 1   1     1    0       0         -40      0                 0      142
## 2   1     1    1       0         -30      0                21      140
## 3   1     1    2       0         -40      0                21      140
## 4   1     1    3       0         -30      0                 0      137
## 5   1     1    4       0         -30      0                 0      133
## 6   1     1    5       0         -28      0                 0      136
##   rh_percent
## 1         84
## 2         71
## 3         84
## 4         77
## 5         77
## 6         79
```

```
summary(df_shasta)
```

```
##       day            month             hour          diffuse      
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  8.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 48.09  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 82.00  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :387.00  
##   temp_tenths        direct       wind_speed_tenths    wind_dir    
##  Min.   :-106.0   Min.   :   0.0   Min.   : 0.00     Min.   :  0.0  
##  1st Qu.:  26.0   1st Qu.:   0.0   1st Qu.: 0.00     1st Qu.: 70.0  
##  Median :  89.0   Median :   0.0   Median : 0.00     Median :172.0  
##  Mean   : 104.4   Mean   : 273.6   Mean   :10.63     Mean   :180.6  
##  3rd Qu.: 172.0   3rd Qu.: 616.0   3rd Qu.:21.00     3rd Qu.:305.2  
##  Max.   : 372.0   Max.   :1008.0   Max.   :67.00     Max.   :360.0  
##    rh_percent    
##  Min.   : 10.00  
##  1st Qu.: 37.00  
##  Median : 56.00  
##  Mean   : 56.21  
##  3rd Qu.: 76.00  
##  Max.   :100.00  
```

```
# 5) Point Arguello
file_path_arguello <- "/Users/chenbohan/Desktop/Solar1/USA_CA_Point.Arguello.994210_TMYx.2009-2023/USA_CA_Point.Arguello.994210_TMYx.2009-2023.clm"
df_arguello <- readClmFile(file_path_arguello)
head(df_arguello)
```

```
##   day month hour diffuse temp_tenths direct wind_speed_tenths wind_dir
## 1   1     1    0       0         119      0                36      360
## 2   1     1    1       0         121      0                51      350
## 3   1     1    2       0         122      0                62       10
## 4   1     1    3       0         124      0                62      360
## 5   1     1    4       0         119      0                51      360
## 6   1     1    5       0         117      0                46      360
##   rh_percent
## 1         85
## 2         84
## 3         81
## 4         80
## 5         79
## 6         80
```

```
summary(df_arguello)
```

```
##       day            month             hour          diffuse      
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  9.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 50.81  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 90.00  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :401.00  
##   temp_tenths        direct      wind_speed_tenths    wind_dir    
##  Min.   : 59.0   Min.   :  0.0   Min.   :  0.00    Min.   :  1.0  
##  1st Qu.:119.0   1st Qu.:  0.0   1st Qu.: 46.00    1st Qu.:281.5  
##  Median :133.0   Median :  0.0   Median : 77.00    Median :350.0  
##  Mean   :134.2   Mean   :297.4   Mean   : 77.71    Mean   :288.8  
##  3rd Qu.:147.0   3rd Qu.:686.2   3rd Qu.:108.00    3rd Qu.:360.0  
##  Max.   :287.0   Max.   :991.0   Max.   :216.00    Max.   :360.0  
##    rh_percent   
##  Min.   :36.00  
##  1st Qu.:75.00  
##  Median :83.00  
##  Mean   :80.97  
##  3rd Qu.:89.00  
##  Max.   :99.00
```

1.Checklist to verify the .clm data frames are correct

1.class()

```
class(df_mammoth)
```

```
## [1] "data.frame"
```

```
class(df_ucdavis)
```

```
## [1] "data.frame"
```

```
class(df_sandiego)
```

```
## [1] "data.frame"
```

```
class(df_shasta)
```

```
## [1] "data.frame"
```

```
class(df_arguello)
```

```
## [1] "data.frame"
```

This confirms that each of your parsed .clm files has been successfully loaded into R as a data frame.

2.dim()

```
dim(df_mammoth)
```

```
## [1] 8760    9
```

```
dim(df_ucdavis)
```

```
## [1] 8760    9
```

```
dim(df_sandiego)
```

```
## [1] 8760    9
```

```
dim(df_shasta)
```

```
## [1] 8760    9
```

```
dim(df_arguello)
```

```
## [1] 8760    9
```

We used dim() to confirm that each location’s data frame has 8760 rows and 9 columns. The 8760 rows match the expected number of hourly observations in a year (365 days × 24 hours), and the 9 columns reflect the correct structure: three time columns (day, month, hour) and six climate variables.

3.names()

```
names(df_mammoth)
```

```
## [1] "day"              "month"            "hour"            
## [4] "diffuse"          "temp_tenths"      "direct"          
## [7] "wind_speed_tenths" "wind_dir"         "rh_percent"
```

```
names(df_ucdavis)
```

```
## [1] "day"              "month"            "hour"            
## [4] "diffuse"          "temp_tenths"      "direct"          
## [7] "wind_speed_tenths" "wind_dir"         "rh_percent"
```

```
names(df_sandiego)
```

```
## [1] "day"              "month"            "hour"            
## [4] "diffuse"          "temp_tenths"      "direct"          
## [7] "wind_speed_tenths" "wind_dir"         "rh_percent"
```

```
names(df_shasta)
```

```
## [1] "day"              "month"            "hour"            
## [4] "diffuse"          "temp_tenths"      "direct"          
## [7] "wind_speed_tenths" "wind_dir"         "rh_percent"
```

```
names(df_arguello)
```

```
## [1] "day"              "month"            "hour"            
## [4] "diffuse"          "temp_tenths"      "direct"          
## [7] "wind_speed_tenths" "wind_dir"         "rh_percent"
```

We used the names() function to verify that each data frame contains the correct column names. All five datasets had identical and expected column labels: day, month, hour, diffuse, temp_tenths, direct, wind_speed_tenths, wind_dir, and rh_percent. This confirms our function accurately extracted and labeled each field from the .clm files.

4.str()

```
str(df_mammoth)
```

```
## 'data.frame':    8760 obs. of  9 variables:
##  $ day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ diffuse         : num  0 0 0 0 0 0 0 0 33 61 ...
##  $ temp_tenths     : num  -90 -90 -90 -90 -80 -110 -120 -130 -90 -50 ...
##  $ direct          : num  0 0 0 0 0 0 0 0 355 774 ...
##  $ wind_speed_tenths: num  26 21 0 0 0 0 0 15 0 21 ...
##  $ wind_dir        : num  10 310 180 187 192 172 174 340 187 310 ...
##  $ rh_percent      : num  77 70 70 64 78 70 69 69 70 54 ...
```

```
str(df_ucdavis)
```

```
## 'data.frame':    8760 obs. of  9 variables:
##  $ day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ diffuse         : num  0 0 0 0 0 0 0 0 23 50 ...
##  $ temp_tenths     : num  50 58 40 39 36 33 30 27 18 41 ...
##  $ direct          : num  0 0 0 0 0 0 0 0 294 676 ...
##  $ wind_speed_tenths: num  31 15 0 21 0 0 21 0 0 21 ...
##  $ wind_dir        : num  30 80 2 120 29 46 150 75 67 140 ...
##  $ rh_percent      : num  42 47 58 56 56 61 66 85 92 61 ...
```

```
str(df_sandiego)
```

```
## 'data.frame':    8760 obs. of  9 variables:
##  $ day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ diffuse         : num  0 110 138 165 200 207 181 171 149 109 ...
##  $ temp_tenths     : num  266 268 269 270 270 270 267 266 267 268 ...
##  $ direct          : num  0 460 576 681 693 661 706 706 586 454 ...
##  $ wind_speed_tenths: num  36 46 41 31 31 41 31 21 0 15 ...
##  $ wind_dir        : num  70 70 80 80 80 70 80 80 169 240 ...
##  $ rh_percent      : num  71 63 60 58 77 62 60 81 81 81 ...
```

```
str(df_shasta)
```

```
## 'data.frame':    8760 obs. of  9 variables:
##  $ day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ diffuse         : num  0 0 0 0 0 0 0 0 17 57 ...
##  $ temp_tenths     : num  -40 -30 -40 -30 -30 -28 -22 -17 -20 -10 ...
##  $ direct          : num  0 0 0 0 0 0 0 0 28 99 ...
##  $ wind_speed_tenths: num  0 21 21 0 0 0 0 0 15 0 ...
##  $ wind_dir        : num  142 140 140 137 133 136 143 143 130 149 ...
##  $ rh_percent      : num  84 71 84 77 77 79 75 76 85 78 ...
```

```
str(df_arguello)
```

```
## 'data.frame':    8760 obs. of  9 variables:
##  $ day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ diffuse         : num  0 0 0 0 0 0 0 0 39 63 ...
##  $ temp_tenths     : num  119 121 122 124 119 117 116 114 114 120 ...
##  $ direct          : num  0 0 0 0 0 0 0 0 279 665 ...
##  $ wind_speed_tenths: num  36 51 62 62 51 46 46 57 67 72 ...
##  $ wind_dir        : num  360 350 10 360 360 360 360 360 360 360 ...
##  $ rh_percent      : num  85 84 81 80 79 80 80 79 77 75 ...
```

We used the str() function to examine the structure of each data frame. Each had 8760 rows and 9 columns, indicating a full year of hourly data. All variables had appropriate data types: integers for day, month, and hour, and numerics for the six climate-related measurements. This confirms our .clm file parsing logic worked consistently and accurately across all locations.

2. Check for Missing Values (NAs)

1.Total NAs

```
sum(is.na(df_mammoth))
```

```
## [1] 0
```

```
sum(is.na(df_ucdavis))
```

```
## [1] 0
```

```
sum(is.na(df_sandiego))
```

```
## [1] 0
```

```
sum(is.na(df_shasta))
```

```
## [1] 0
```

```
sum(is.na(df_arguello))
```

```
## [1] 0
```

We programmatically verified the completeness of each dataset by checking for missing values using sum(is.na(…)). All five locations — Mammoth, UC Davis, San Diego, Mount Shasta, and Point Arguello — returned zero missing values, confirming that all rows and columns were successfully read and parsed with no data loss.

2.NAs per Column

```
#NAs per Column
sapply(df_mammoth, function(col) sum(is.na(col)))
```

```
##               day             month              hour           diffuse 
##                 0                 0                 0                 0 
##       temp_tenths            direct wind_speed_tenths          wind_dir 
##                 0                 0                 0                 0 
##        rh_percent 
##                 0
```

```
sapply(df_ucdavis, function(col) sum(is.na(col)))
```

```
##               day             month              hour           diffuse 
##                 0                 0                 0                 0 
##       temp_tenths            direct wind_speed_tenths          wind_dir 
##                 0                 0                 0                 0 
##        rh_percent 
##                 0
```

```
sapply(df_sandiego, function(col) sum(is.na(col)))
```

```
##               day             month              hour           diffuse 
##                 0                 0                 0                 0 
##       temp_tenths            direct wind_speed_tenths          wind_dir 
##                 0                 0                 0                 0 
##        rh_percent 
##                 0
```

```
sapply(df_shasta, function(col) sum(is.na(col)))
```

```
##               day             month              hour           diffuse 
##                 0                 0                 0                 0 
##       temp_tenths            direct wind_speed_tenths          wind_dir 
##                 0                 0                 0                 0 
##        rh_percent 
##                 0
```

```
sapply(df_arguello, function(col) sum(is.na(col)))
```

```
##               day             month              hour           diffuse 
##                 0                 0                 0                 0 
##       temp_tenths            direct wind_speed_tenths          wind_dir 
##                 0                 0                 0                 0 
##        rh_percent 
##                 0
```

We verified that each individual column in the data frames contained no missing values. We used sapply(…, function(col) sum(is.na(col))) to check each variable. All columns across all five locations — including diffuse, temp_tenths, direct, and others — returned zero NAs, confirming the integrity and completeness of the imported data at the column level.

3.Use summary() to get min, max, median, mean, etc

```
#Use summary() to get min, max, median, mean, etc
summary(df_mammoth)
```

```
##       day            month             hour          diffuse
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00
##  Median :16.00   Median : 7.000   Median :11.50   Median : 10.00
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 53.45
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 86.00
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :512.00
##   temp_tenths          direct       wind_speed_tenths    wind_dir
##  Min.   :-248.00   Min.   :   0.0   Min.   :  0.00    Min.   :  0.0
##  1st Qu.:   4.00   1st Qu.:   0.0   1st Qu.:  0.00    1st Qu.:165.0
##  Median :  51.00   Median :   0.0   Median : 26.00    Median :227.0
##  Mean   :  67.21   Mean   : 317.5   Mean   : 32.05    Mean   :220.3
##  3rd Qu.: 130.00   3rd Qu.: 743.0   3rd Qu.: 46.00    3rd Qu.:280.0
##  Max.   : 350.00   Max.   :1080.0   Max.   :227.00    Max.   :360.0
##    rh_percent
##  Min.   : 10.00
##  1st Qu.: 34.00
##  Median : 51.00
##  Mean   : 51.14
##  3rd Qu.: 68.00
##  Max.   :100.00
```

```
summary(df_ucdavis)
```

```
##       day            month             hour          diffuse      
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  9.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 47.86  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 88.25  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :432.00  
##   temp_tenths        direct      wind_speed_tenths    wind_dir    
##  Min.   :-10.0   Min.   :  0.0   Min.   :  0.00    Min.   :  0.0  
##  1st Qu.:110.0   1st Qu.:  0.0   1st Qu.:  6.00    1st Qu.:170.0  
##  Median :154.0   Median :  0.0   Median : 21.00    Median :210.0  
##  Mean   :168.3   Mean   :282.4   Mean   : 24.05    Mean   :215.9  
##  3rd Qu.:220.0   3rd Qu.:667.0   3rd Qu.: 36.00    3rd Qu.:292.0  
##  Max.   :390.0   Max.   :994.0   Max.   :129.00    Max.   :360.0  
##    rh_percent    
##  Min.   : 10.00  
##  1st Qu.: 37.00  
##  Median : 59.00  
##  Mean   : 58.01  
##  3rd Qu.: 78.00  
##  Max.   :100.00  
```

```
summary(df_sandiego)
```

```
##       day            month             hour          diffuse      
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  3.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 75.12  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.:151.25  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :429.00  
##   temp_tenths        direct      wind_speed_tenths    wind_dir    
##  Min.   :172.0   Min.   :  0.0   Min.   :  0.00    Min.   : 10.0  
##  1st Qu.:234.0   1st Qu.:  0.0   1st Qu.: 15.00    1st Qu.:147.0  
##  Median :258.0   Median :  0.0   Median : 26.00    Median :208.0  
##  Mean   :250.1   Mean   :194.4   Mean   : 25.24    Mean   :206.9  
##  3rd Qu.:271.0   3rd Qu.:453.2   3rd Qu.: 36.00    3rd Qu.:290.0  
##  Max.   :322.0   Max.   :891.0   Max.   :113.00    Max.   :360.0  
##    rh_percent    
##  Min.   : 54.00  
##  1st Qu.: 74.00  
##  Median : 80.00  
##  Mean   : 79.07  
##  3rd Qu.: 84.00  
##  Max.   :100.00  
```

```
summary(df_shasta)
```

```
##       day            month             hour          diffuse     
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  8.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 48.09  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 82.00  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :387.00  
##   temp_tenths         direct       wind_speed_tenths    wind_dir    
##  Min.   :-106.0   Min.   :   0.0   Min.   : 0.00     Min.   :  0.0  
##  1st Qu.:  26.0   1st Qu.:   0.0   1st Qu.: 0.00     1st Qu.: 70.0  
##  Median :  89.0   Median :   0.0   Median : 0.00     Median :172.0  
##  Mean   : 104.4   Mean   : 273.6   Mean   :10.63     Mean   :180.6  
##  3rd Qu.: 172.0   3rd Qu.: 616.0   3rd Qu.:21.00     3rd Qu.:305.2  
##  Max.   : 372.0   Max.   :1008.0   Max.   :67.00     Max.   :360.0  
##    rh_percent    
##  Min.   : 10.00  
##  1st Qu.: 37.00  
##  Median : 56.00  
##  Mean   : 56.21  
##  3rd Qu.: 76.00  
##  Max.   :100.00
```

```
summary(df_arguello)
```

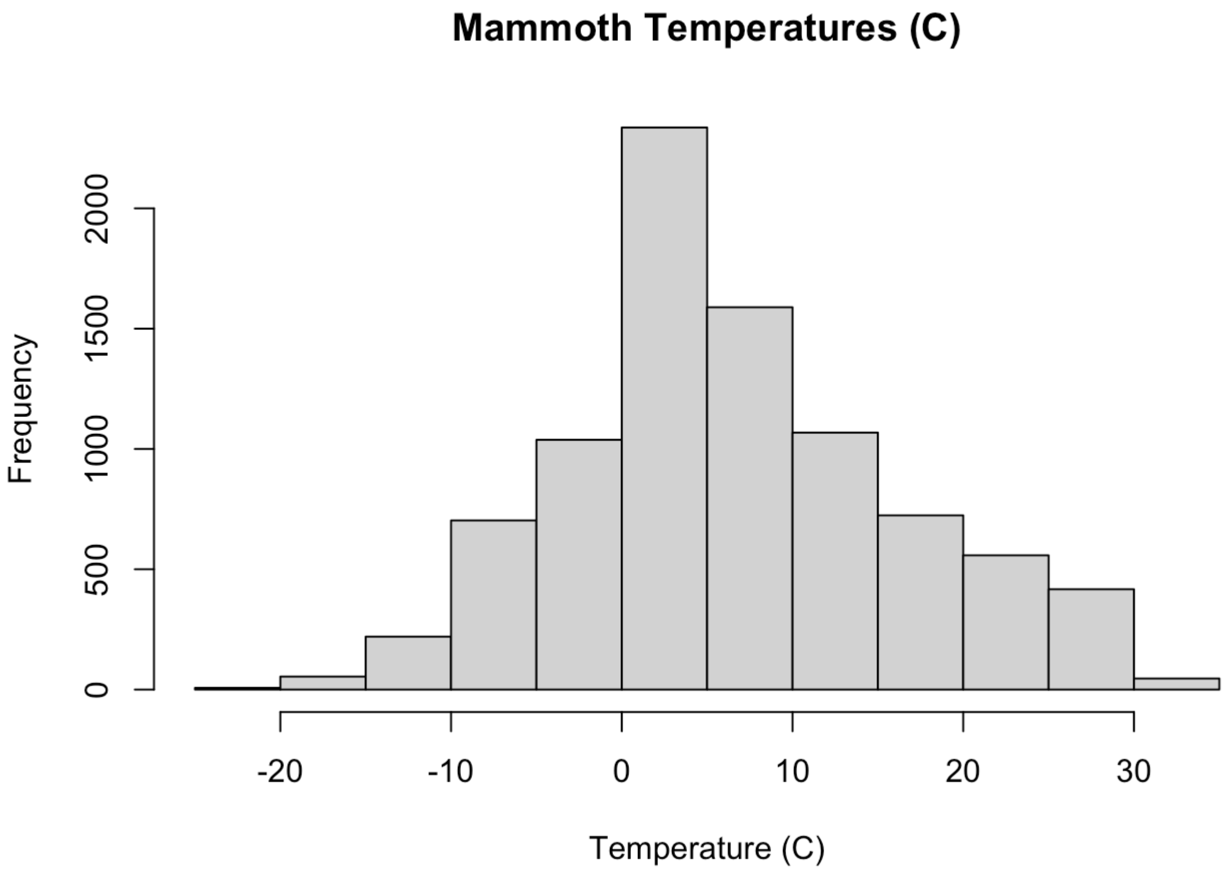
```
##       day            month             hour          diffuse     
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.00   Min.   :  0.00  
##  1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.: 5.75   1st Qu.:  0.00  
##  Median :16.00   Median : 7.000   Median :11.50   Median :  9.00  
##  Mean   :15.72   Mean   : 6.526   Mean   :11.50   Mean   : 50.81  
##  3rd Qu.:23.00   3rd Qu.:10.000   3rd Qu.:17.25   3rd Qu.: 90.00  
##  Max.   :31.00   Max.   :12.000   Max.   :23.00   Max.   :401.00  
##   temp_tenths        direct      wind_speed_tenths    wind_dir    
##  Min.   : 59.0   Min.   :  0.0   Min.   :  0.00    Min.   :  1.0  
##  1st Qu.:119.0   1st Qu.:  0.0   1st Qu.: 46.00    1st Qu.:281.5  
##  Median :133.0   Median :  0.0   Median : 77.00    Median :350.0  
##  Mean   :134.2   Mean   :297.4   Mean   : 77.71    Mean   :288.8  
##  3rd Qu.:147.0   3rd Qu.:686.2   3rd Qu.:108.00    3rd Qu.:360.0  
##  Max.   :287.0   Max.   :991.0   Max.   :216.00    Max.   :360.0  
##    rh_percent   
##  Min.   :36.00  
##  1st Qu.:75.00  
##  Median :83.00  
##  Mean   :80.97  
##  3rd Qu.:89.00  
##  Max.   :99.00
```

We used the summary() function to inspect the range and distribution of each variable across all five datasets. This confirmed several key assumptions: That day ranged from 1 to 31, month from 1 to 12, and hour from 0 to 23, which supports the assumption that each day is represented by exactly 24 rows. That all climate measurements (e.g., temperature in tenths of °C, solar radiation in W/m², humidity in %) were within plausible physical ranges. That no obviously invalid values appeared.

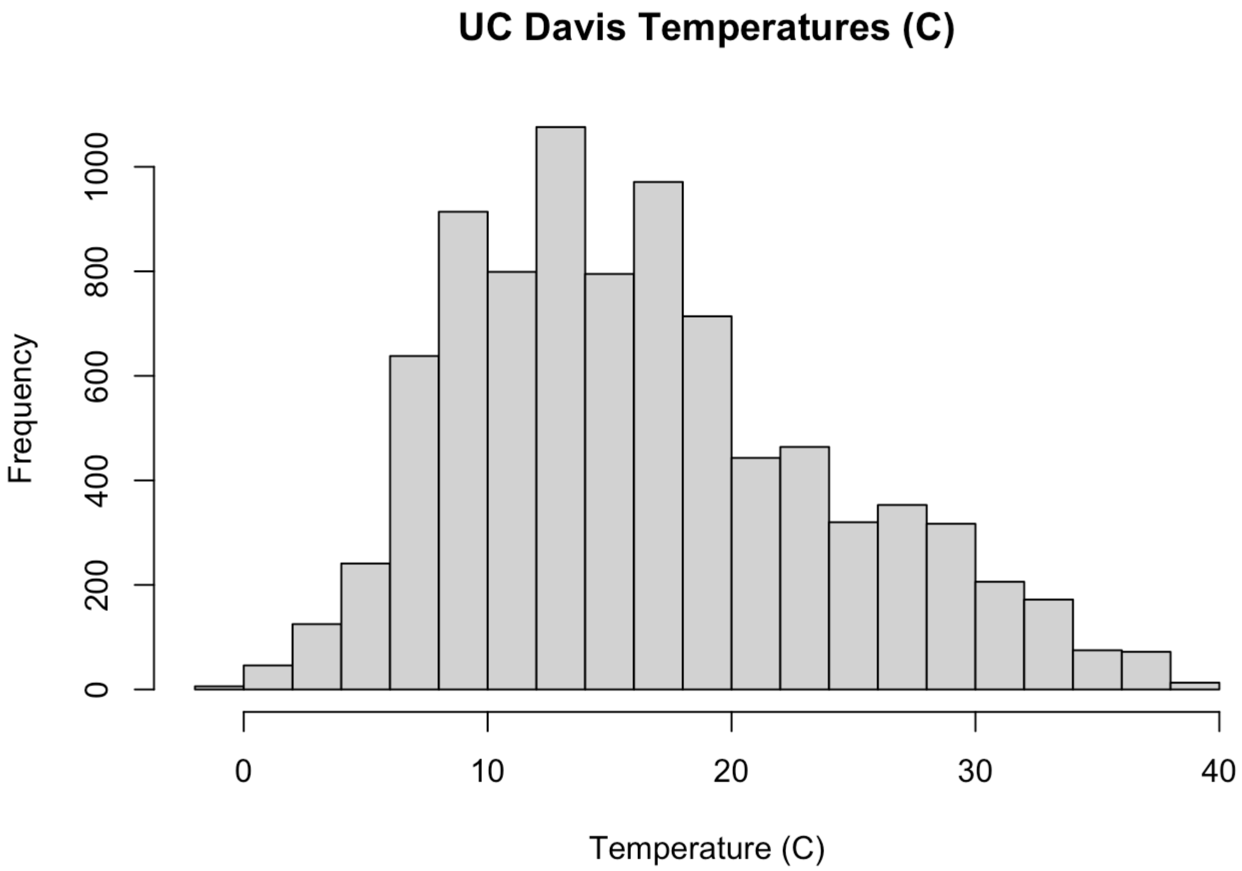
4.Graphical Checks

1.Histograms

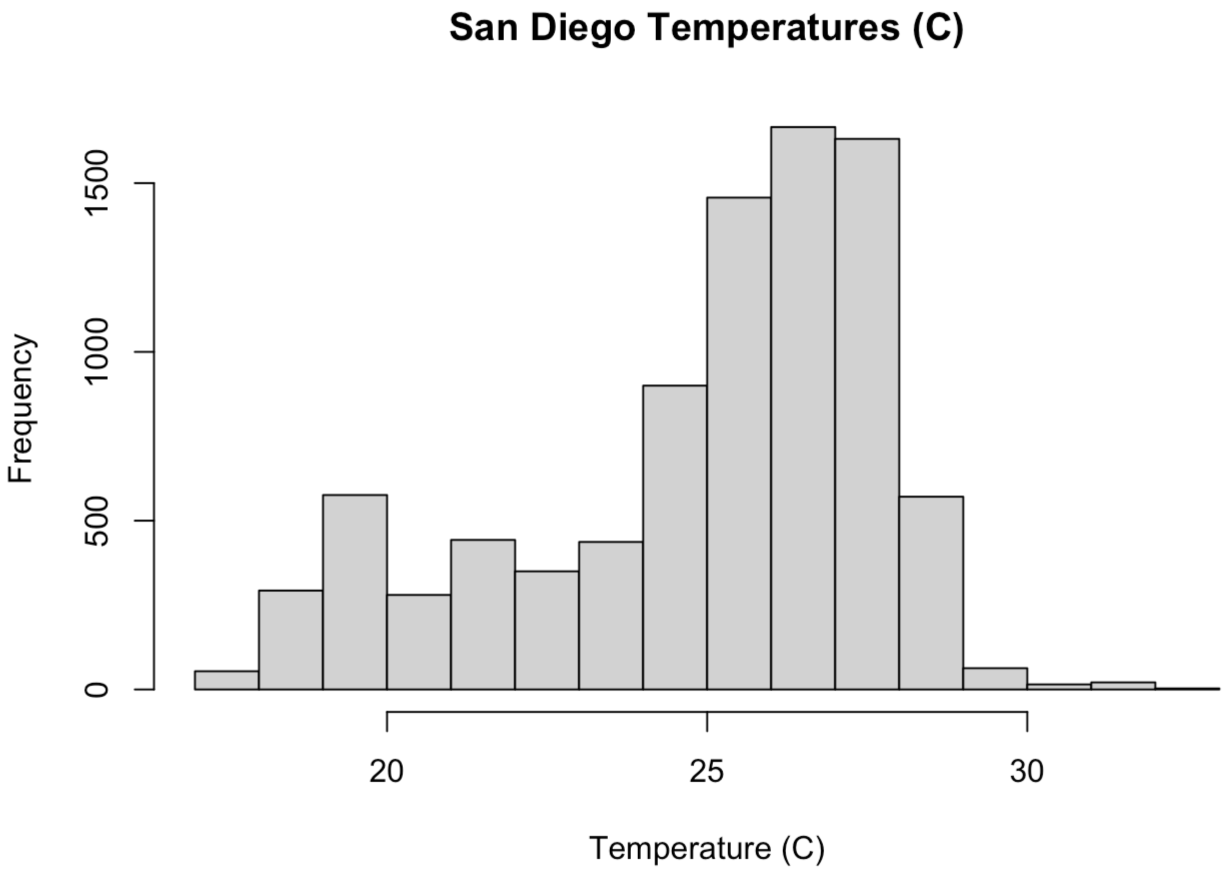
```
#Histograms
hist(df_mammoth$temp_tenths / 10, main = "Mammoth Temperatures (C)",
     xlab = "Temperature (C)")
```



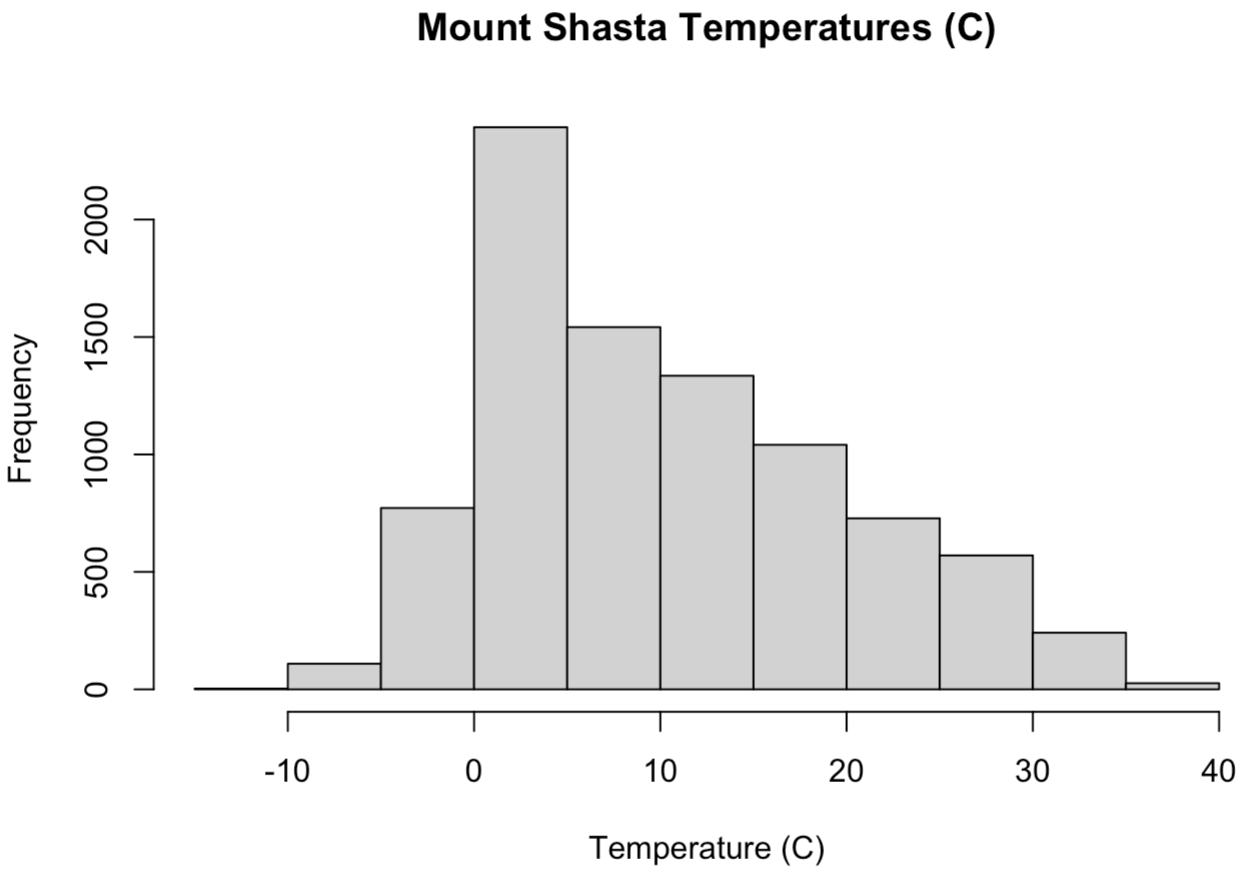
```
hist(df_ucdavis$temp_tenths / 10,
     main = "UC Davis Temperatures (C)",
     xlab = "Temperature (C)")
```



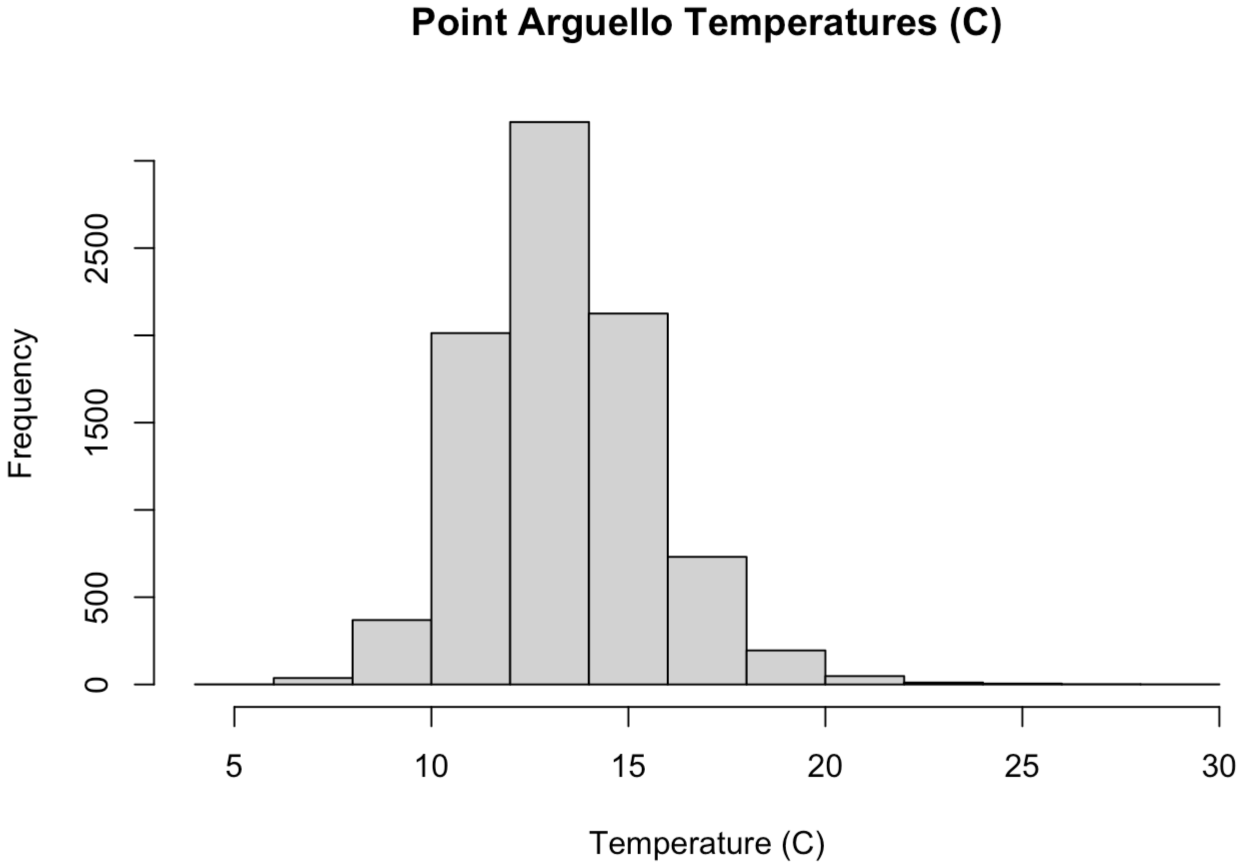
```
hist(df_sandiego$temp_tenths / 10,
     main = "San Diego Temperatures (C)",
     xlab = "Temperature (C)")
```



```
hist(df_shasta$temp_tenths / 10,
     main = "Mount Shasta Temperatures (C)",
     xlab = "Temperature (C)")
```



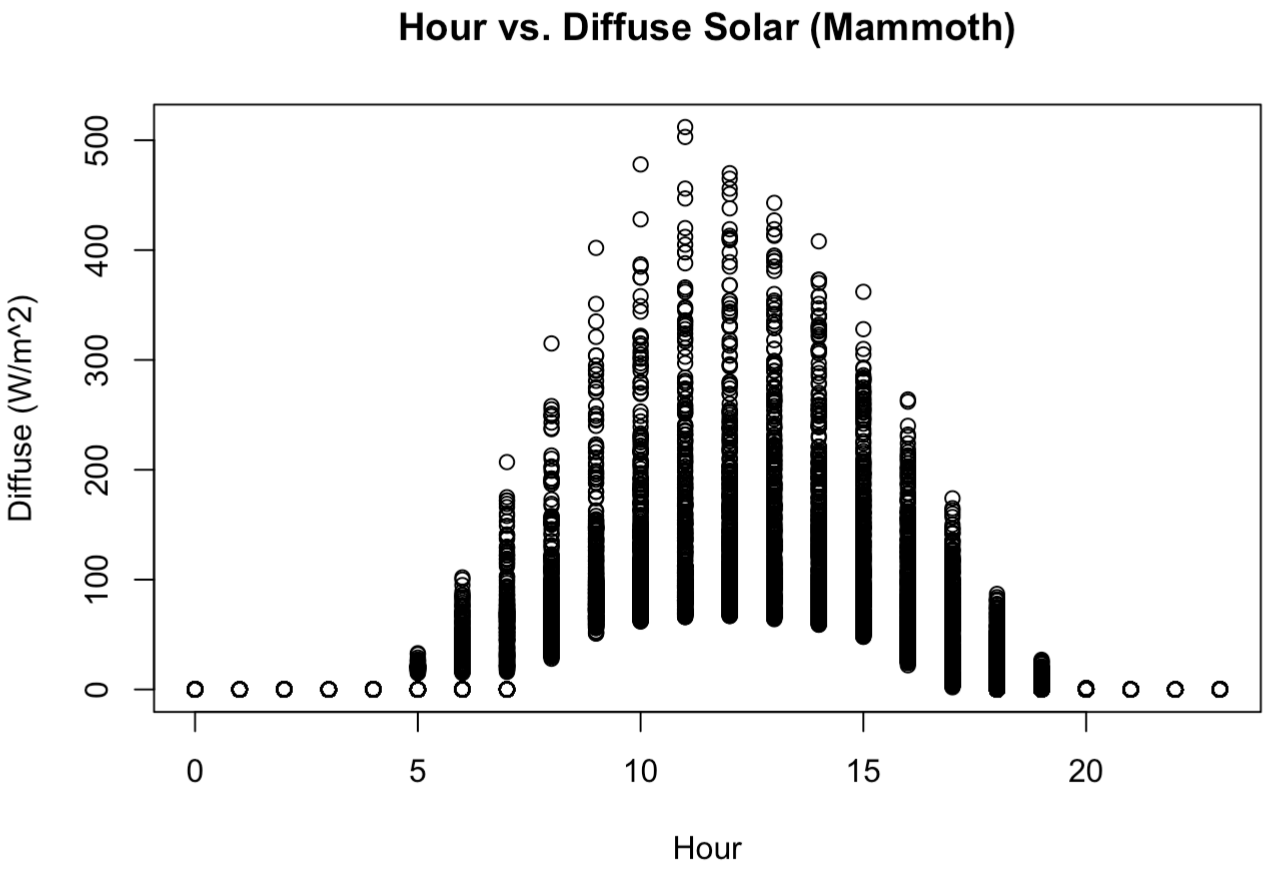
```
hist(df_arguello$temp_tenths / 10,
     main = "Point Arguello Temperatures (C)",
     xlab = "Temperature (C)")
```



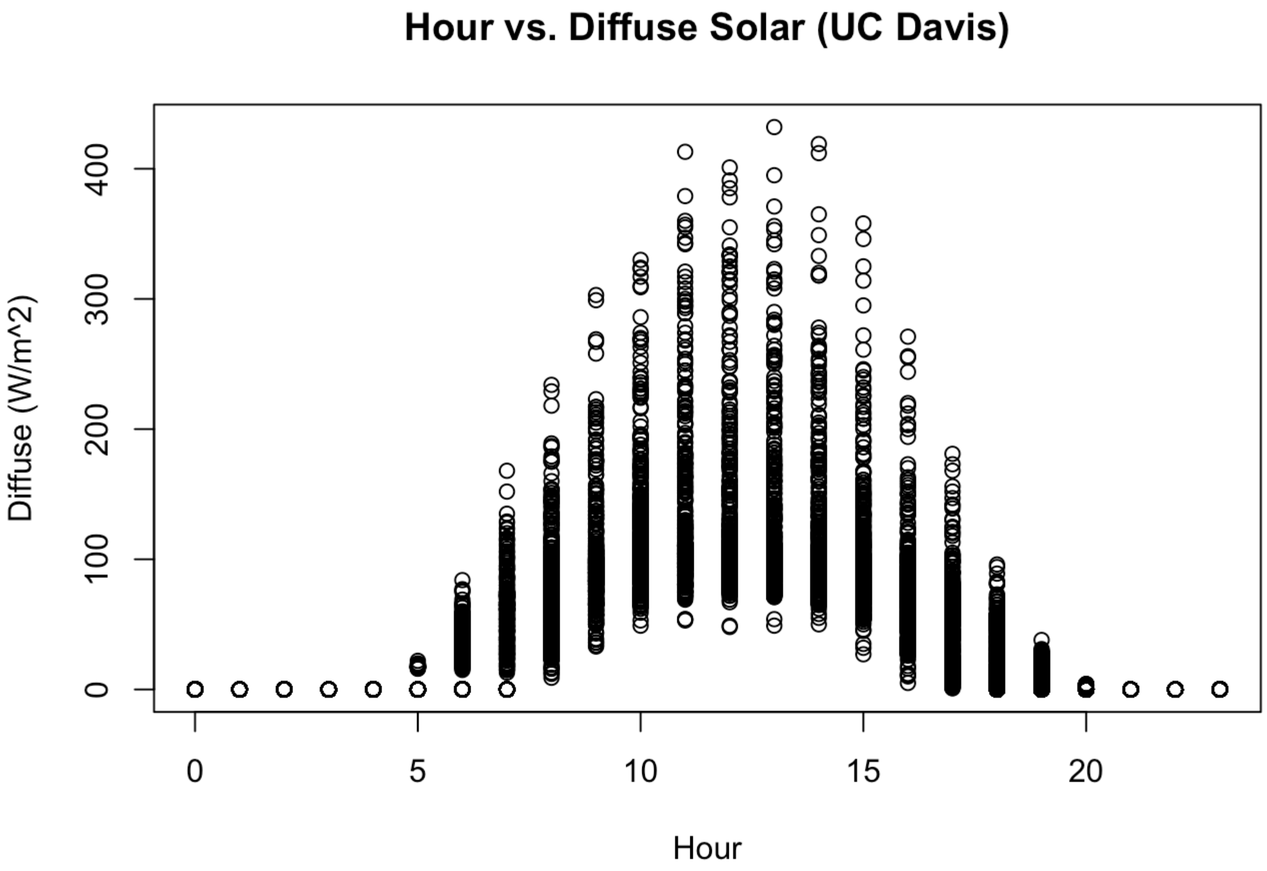
These plots help verify: Units were handled correctly (i.e., converted from tenths of °C). No outliers or unrealistic values appeared (e.g., no temperatures >50°C or <-30°C). The shapes and ranges of the distributions match our expectations based on each location's real-world climate. No parsing errors or missing data blocks

2.Scatter Plots

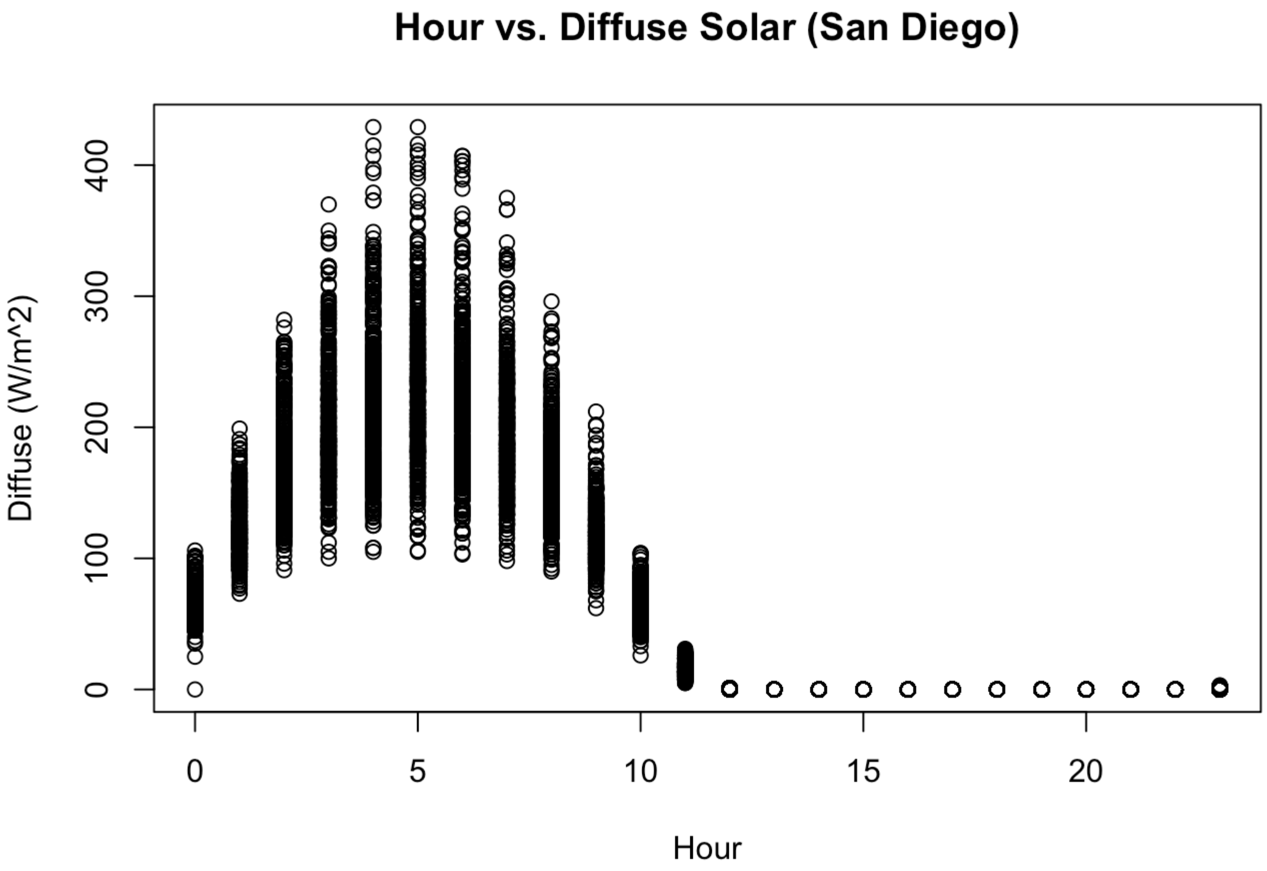
```
#Scatter Plots
plot(df_mammoth$hour, df_mammoth$diffuse,
     main = "Hour vs. Diffuse Solar (Mammoth)",
     xlab = "Hour", ylab = "Diffuse (W/m^2)")
```



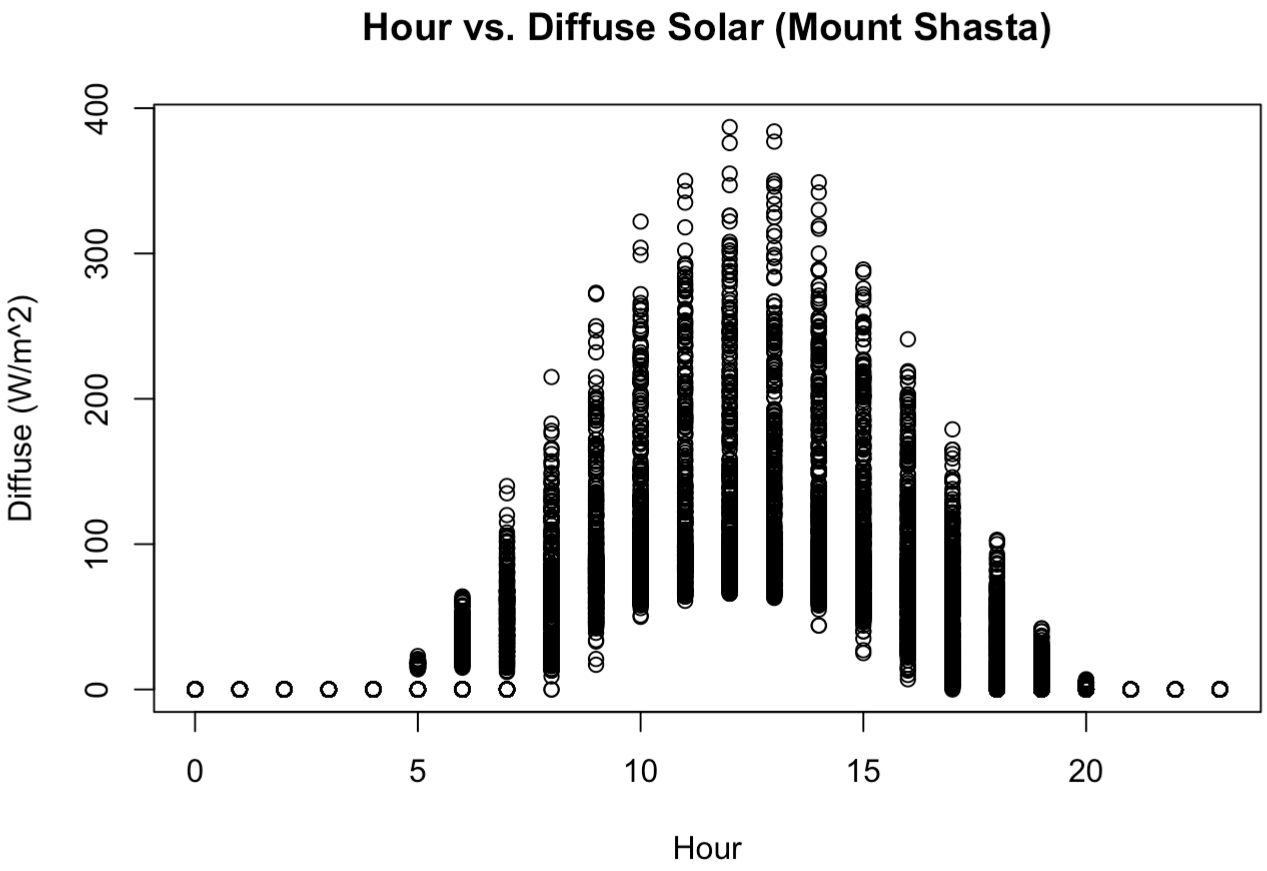
```
plot(df_ucdavis$hour, df_ucdavis$diffuse,
     main = "Hour vs. Diffuse Solar (UC Davis)",
     xlab = "Hour", ylab = "Diffuse (W/m^2)")
```



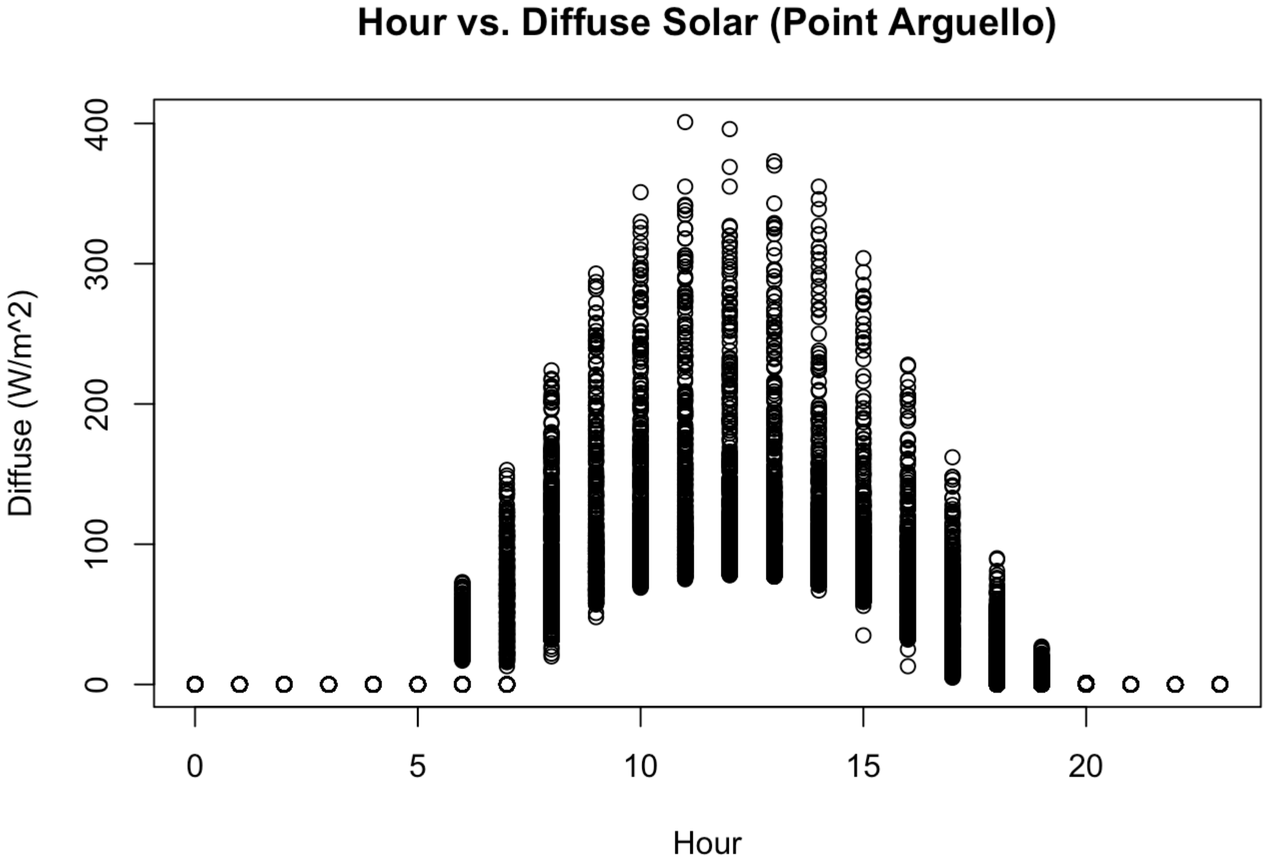
```
plot(df_sandiego$hour, df_sandiego$diffuse,
     main = "Hour vs. Diffuse Solar (San Diego)",
     xlab = "Hour", ylab = "Diffuse (W/m^2)")
```



```
plot(df_shasta$hour, df_shasta$diffuse,
     main = "Hour vs. Diffuse Solar (Mount Shasta)",
     xlab = "Hour", ylab = "Diffuse (W/m^2)")
```



```
plot(df_arguello$hour, df_arguello$diffuse,
     main = "Hour vs. Diffuse Solar (Point Arguello)",
     xlab = "Hour", ylab = "Diffuse (W/m^2)")
```



These plots confirm that the diffuse solar radiation values follow expected daily patterns, giving strong evidence that:

The .clm files were read and parsed correctly. The hour column is accurate. The diffuse column is interpreted correctly (units and scaling). The datasets are complete for 24-hour cycles

Manually comparing individual values in the files and the results

I Open the .clm file manually navigate to:/Users/chenbohan/Desktop/Solar1/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2023/ And open it with Textedit and then I locate the target lines:* day 1 month 1 and the data is below: * day 1 month 1 0,-90,0,26,10,77 0,-90,0,21,310,70 0,-90,0,0,180,70 0,-90,0,0,187,64 0,-80,0,0,192,78 0,-110,0,0,172,70 0,-120,0,0,174,69 0,-130,0,15,340,69 33,-90,355,0,187,70 61,-50,774,21,310,54 73,-10,897,21,130,42 79,10,967,31,140,33 81,50,975,21,120,44 83,60,944,21,130,38 87,8,834,26,110,37 88,4,554,15,120,40 53,5,0,31,110,37 10,0,0,0,208,39 0,-20,0,15,150,46 0,-40,0,15,60,59 0,-70,0,0,221,70 0,-90,0,15,30,57 0,-90,0,0,215,58 0,-90,0,21,250,60 I compare the data with the corresponding rows in df_mammoth. The values matched exactly for all six columns, confirming that the parsing function extracted the correct data. And then I do both steps for other 4 locations and get the same result which confirms that the parsing function extracted the correct data for both 5 locations.

Identify Assumptions

Assumption: The files represent data across a full year with months ranging from 1 to 12 for all 5 locations. Verification: We summary() on each data frame and observed that the month variable spans from 1 to 12 in all cases, as expected.

Assumption: Each location’s .clm file should contain 365 days × 24 hours = 8,760 rows of data. Verification: We used dim(df_) for each data frame and confirmed that each has exactly 8,760 rows and 9 columns

Assumption: Day values should be between 1 and 31 depending on the month. Verification: We summary() on each data frame and observed that day ranges from 1–31, consistent with expectations.

ANOVA Table

In this study, a one-way ANOVA was applied to compare daily average temperature, daily average relative humidity, daily average wind speed, and total daily solar radiation across five different California weather stations. This test allows us to formally evaluate whether location has a statistically significant effect on these key climate variables.

```
library(dplyr)
```

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

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(readr)

# Five location path
files <- c(
"/Users/chenbohan/Desktop/Solar1/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2023 2/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2023.clm",
"/Users/chenbohan/Desktop/Solar1/USA_CA_Mount.Shasta.725957_TMYx.2009-2023/USA_CA_Mount.Shasta.725957_TMYx.2009-2023.clm",
"/Users/chenbohan/Desktop/Solar1/USA_CA_Point.Arguello.994210_TMYx.2009-2023/USA_CA_Point.Arguello.994210_TMYx.2009-2023.clm",
"/Users/chenbohan/Desktop/Solar1/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.2009-2023/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.2009-2023.clm",
"/Users/chenbohan/Desktop/Solar1/USA_CA_UC-Davis-University.AP.720576_TMYx.2009-2023/USA_CA_UC-Davis-University.AP.720576_TMYx.2009-2023.clm"
)
files
```

```
## [1] "/Users/chenbohan/Desktop/Solar1/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2
023 2/USA_CA_Mammoth.Yosemite.AP.723894_TMYx.2009-2023.clm"
## [2] "/Users/chenbohan/Desktop/Solar1/USA_CA_Mount.Shasta.725957_TMYx.2009-2023/USA
_CA_Mount.Shasta.725957_TMYx.2009-2023.clm"
## [3] "/Users/chenbohan/Desktop/Solar1/USA_CA_Point.Arguello.994210_TMYx.2009-2023/U
SA_CA_Point.Arguello.994210_TMYx.2009-2023.clm"
## [4] "/Users/chenbohan/Desktop/Solar1/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.200
9-2023/USA_CA_San.Diego-MCAS.Miramar.722930_TMYx.2009-2023.clm"
## [5] "/Users/chenbohan/Desktop/Solar1/USA_CA_UC-Davis-University.AP.720576_TMYx.200
9-2023/USA_CA_UC-Davis-University.AP.720576_TMYx.2009-2023.clm"
```

```
all_data <- bind_rows(lapply(files, function(f){
  df <- readClmFile(f)
  df$location <- sub("_TMYx.*","", basename(dirname(f)))
  return(df)
}))

daily_data <- all_data %>%
  mutate(
    temp = temp_tenths / 10,
    wind = wind_speed_tenths / 10,
    solar = diffuse + direct
  ) %>%
  group_by(location, month, day) %>%
  summarise(
    mean_temp = mean(temp),
    mean_rh   = mean(rh_percent),
    mean_wind = mean(wind),
    total_solar = sum(solar),
    .groups = "drop"
  )
anova_temp  <- summary(aov(mean_temp ~ location, data = daily_data))
anova_rh    <- summary(aov(mean_rh ~ location, data = daily_data))
anova_wind  <- summary(aov(mean_wind ~ location, data = daily_data))
anova_solar <- summary(aov(total_solar ~ location, data = daily_data))

print(anova_temp)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## location       4  70827   17707   496.1 <2e-16 ***
## Residuals   1820  64965      36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(anova_rh)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## location       4 281365   70341   355.9 <2e-16 ***
## Residuals   1820 359665     198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(anova_wind)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## location       4   9621  2405.3     723 <2e-16 ***
## Residuals   1820   6055     3.3                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(anova_solar)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)    
## location       4 1.202e+09 300493560   49.41 <2e-16 ***
## Residuals   1820 1.107e+10   6081084                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

For all four ANOVA tests (daily average temperature, daily average relative humidity, daily average wind speed, and daily total solar radiation), the p-values are all less than 2 × 10^-16, which is far below the significance level α = 0.05. Therefore, we reject the null hypothesis that the mean values of these climate variables are equal across the five locations. In other words, we conclude that there are statistically significant differences in daily temperature, humidity, wind speed, and solar radiation among the five California weather stations. This result suggests that geographic location has a strong and significant impact on regional climate characteristics.