CMI Data Analysis

Jonathan Kang (jk36)

2022-10-17

```
library(tidyverse)
library(readr)
library(sqldf)
library(lubridate)
library(ggplot2)
library(plotly)
library(weathermetrics)
```

Import .txt data files

(Refer to exploratory for details)

```
# original
ogCMIDAY <- read.delim("~/Documents/College/FA22/URES/RawData/CMIDAY.txt")
units <- head(ogCMIDAY,1) # units for each category

# using bash, remove the second row with the units, create new file without the units
  # sed 2d CMIDAY.txt > CMIDAY_rmunits.txt
CMIDAY_rmunits <- read.delim("~/Documents/College/FA22/URES/RawData/CMIDAY_rmunits.txt")
# glimpse(CMIDAY_rmunits)

CMIDAY_data <- head(CMIDAY_rmunits, -10) # excess notes at bottom
# head(CMIDAY_data)

data <- sqldf("
SELECT year, month, day, avg_rel_hum, avg_air_temp as avg_air_temp_f
FROM CMIDAY_data
")

data$avg_rel_hum <- as.numeric(data$avg_rel_hum)
data$avg_air_temp_f <- as.numeric(data$avg_air_temp_f)

# Fix the NA issues, then proceed with Celsius version and create time series for that
data2 = data

mean_temp1 = mean(c(data$avg_air_temp_f[(8325-6):8325],data$avg_air_temp_f[8328:(8328+6)]))
mean_hum1 = mean(c(data$avg_rel_hum[(8325-6):8325],data$avg_rel_hum[8328:(8328+6)]))

mean_temp2 = mean(c(data$avg_air_temp_f[(9212-6):9212],data$avg_air_temp_f[9214:(9214+6)]))
mean_hum2 = mean(c(data$avg_rel_hum[(9212-6):9212],data$avg_rel_hum[9214:(9214+6)]))
```

```r
mean_temp3 = mean(c(data$avg_air_temp_f[(9990-6):9990],data$avg_air_temp_f[9994:(9994+6)]))
mean_hum3 = mean(c(data$avg_rel_hum[(9990-6):9990],data$avg_rel_hum[9994:(9994+6)]))
data2$avg_air_temp_f[8326:8327] = round(mean_temp1,3)
data2$avg_air_temp_f[9213] = round(mean_temp2,3)
data2$avg_air_temp_f[9991:9993] = round(mean_temp3,3)
data2$avg_rel_hum[8326:8327] = round(mean_hum1,3)
data2$avg_rel_hum[9213] = round(mean_hum2,3)
data2$avg_rel_hum[9991:9993] = round(mean_hum3,3)
data2 = data2[-11763,] # remove empty row

data_Xna <- data2 %>% mutate(
  avg_air_temp_c = (avg_air_temp_f - 32) * (5/9),
  date_string = paste(year, month, day, sep="-"),
  date = as.Date(date_string)
)

data_HI <- sqldf("
SELECT date, avg_air_temp_f, avg_rel_hum, month
FROM data_Xna
")

# Bound of avg_rel_hum = [0,100] (only have vals > 100%)
large_rh = grep(T, data_HI$avg_rel_hum > 100)
# slice(data_HI, large_rh)
data_HI$avg_rel_hum[large_rh] = 100

data_HI$heat_index <- heat.index(t = data_HI$avg_air_temp_f,
                                 rh = data_HI$avg_rel_hum,
                                 temperature.metric = "fahrenheit")

# the 85th percentile
heat_index_calc <- data_HI %>%
  filter(month == "7" | month == "8") %>%
  filter(date >= "1990-01-01" & date <"2021-01-01")

HI_index = quantile(heat_index_calc$heat_index, 0.85)
# HI_index
```

```r
data_HW_setup <- data_HI %>%
  mutate(heat_index_85p = heat_index >= HI_index) %>% # heat_index_85p = T if heat_index > 85th of summ
  select(date, heat_index, heat_index_85p) %>%
  filter(month(date) < 10 & month(date) > 4) %>%  # within the interested range
  filter(year(date) >= 1989 & year(date) <= 2020)
  # filter(heat_index_85p == T)

# dates = data_HW_setup$date
# sect = rep(-1,length(dates))
# sect_index = 0
# sect[1] = sect_index
# for (day_index in (2:(length(dates)))) {
#   curr = dates[day_index]
#   yest = dates[day_index - 1]
#   tmrw = dates[day_index + 1]
```

```
#
#   if(curr == (yest + 1)) { # if consec, add to section
#     sect[day_index] = sect_index
#   }
#   else if(curr == (tmrw - 1)) { # if new consecutive, create and add new section
#     sect_index = sect_index + 1
#     sect[day_index] = sect_index
#   }
# }
#
# data_HW_setup$section <- sect
# data_HW <- data_HW_setup %>% filter(section >= 0) %>%
#   select(date, heat_index, section)
```

Intro Analysis

daily

Means = all heat indexes (not limited to Heat Waves)

```
data_HW <- data_HW_setup
attach(data_HW)
```

```
mu = mean(heat_index)
sigmax = sd(heat_index)

lags = c(1,2,7)
covars = c()
correlations = c()

for(lag in lags){
  curr = c()
  lagx = c()
  for(yr in 1:31){ # every year
    for(i in 1:(153 - lag)){ # days in a year (lag is end day)
      curr = c(curr,heat_index[i + yr*153])
      lagx = c(lagx,heat_index[i + lag + yr*153])
      # print(date[i + yr*153])
      # numer = c(numer, (curr - xbar) * (lagx - xbar))
    }
  }

  xbar = mean(curr)
  sx = sd(curr) # sqrt(sum((curr - xbar)**2) / length(curr))
  ybar = mean(lagx)
  sy = sd(lagx)

  # correlation = sum(curr - xbar) * sum(lagx - ybar) / (sx*sy)
  # correlation
  covars = c(covars, cov(curr, lagx))
  correlations = c(correlations, (cov(curr, lagx) / (sx*sy)))
```

```
}
lags
```

```
## [1] 1 2 7
```

```
covars
```

```
## [1] 72.59063 56.61109 32.93637
```

```
correlations
```

```
## [1] 0.8475820 0.6681805 0.4033878
```