clean_data

Viet Dao

6/15/2020

Hanoi

2012-09-04 - 2013-08-30

```
hanoi_orig <- read.csv('./Data/hanoi.csv', header = TRUE, stringsAsFactors = FALSE)
keepCols_hanoi <- c('NAME', 'DATE', 'PRCP', 'TAVG', 'TMAX', 'TMIN')
hanoi <- hanoi_orig[, keepCols_hanoi]

# no of NAs in each row
# sapply(hanoi, function(x) sum(is.na(x)));

# calculate NA TMIN and TMAX from TAVG and each other
hanoi[is.na(hanoi$TMIN) & !is.na(hanoi$TMAX), 'TMIN'] <- hanoi[is.na(hanoi$TMIN) & !is.na(hanoi$TMAX),
hanoi[!is.na(hanoi$TMIN) & is.na(hanoi$TMAX), 'TMAX'] <- hanoi[!is.na(hanoi$TMIN) & is.na(hanoi$TMAX),
# 3 rows: both TMAX and TMIN are NA, replace by TAVG
hanoi[is.na(hanoi$TMIN) & is.na(hanoi$TMAX), 'TMIN'] <- hanoi[is.na(hanoi$TMIN) & is.na(hanoi$TMAX), 'TA
hanoi[is.na(hanoi$TMAX), 'TMAX'] <- hanoi[is.na(hanoi$TMAX), 'TAVG']

# replace 1 NA PRCP by 0
hanoi['PRCP'][is.na(hanoi['PRCP'])] <- 0.0

# set DATE to Date object
hanoi$DATE <- as.Date(hanoi$DATE)
hanoi <- hanoi[order(hanoi$DATE),]

# change name from 'HA DONG' to 'HANOI' for simplicity
hanoi$NAME <- 'HANOI'

# add SNOW and SNWD columns
hanoi$SNOW <- 0
hanoi$SNWD <- 0

# dont need TAVG
hanoi <- hanoi[, !(names(hanoi) %in% ('TAVG'))]
hanoi <- hanoi[, c('NAME', 'DATE', 'PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')]
```

St. Peter

2013-08-15 - 2017-05-31

```
stpeter_orig <- read.csv('./Data/stpeter.csv', header = TRUE, stringsAsFactors = FALSE)
keepCols_stpeter <- c('NAME', 'DATE', 'PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')
stpeter <- stpeter_orig[, keepCols_stpeter];
# stpeter %>% group_by(NAME) %>% summarise(n = n())

# set DATE to Date object
stpeter$DATE <- as.Date(stpeter$DATE)
stpeter <- stpeter[order(stpeter$DATE),]

# fill NAs
stpeter <- aggregate(stpeter, by=list(DATE_ID=stpeter$DATE), min, na.rm = TRUE)
stpeter <- stpeter[, !(names(stpeter) %in% ('DATE_ID'))]

# sapply(stpeter, function(x) sum(is.infinite(x)))

# few Inf is okay
View(stpeter %>% filter(is.infinite(PRCP)|is.infinite(SNWD)|is.infinite(TMAX)|is.infinite(TMIN)))

# rename for simplicty
stpeter$NAME <- 'STPETER'
```

San Francisco

2017-06-01 - 2019-09-05

```
sf_orig <- read.csv('./Data/sf.csv', header = TRUE, stringsAsFactors = FALSE);
keepCols_sf <- c('NAME', 'DATE', 'PRCP', 'SNOW', 'SNWD', 'TMAX', 'TMIN')
sf <- sf_orig[, keepCols_sf]

sapply(sf, function(x) sum(is.na(x)))
```

```
##  NAME  DATE  PRCP  SNOW  SNWD  TMAX  TMIN
##     0     0  8635 36296 56529 37259 37280
```

```
sf$DATE <- as.Date(sf$DATE)
sf <- sf[order(sf$DATE),]

sf <- sf %>% filter(NAME == 'SAN FRANCISCO DOWNTOWN, CA US')
sf$NAME <- 'SF'

sf[c('SNOW', 'SNWD')][is.na(sf[c('SNOW', 'SNWD')])] <- 0
```

Oakland

2018-05-01 - 2019-09-05

```
oakland_orig <- read.csv('./Data/oakland.csv', header = TRUE, stringsAsFactors = FALSE)
oakland <- oakland_orig[, keepCols_sf]
sapply(oakland, function(x) sum(is.na(x)))
```

```
##  NAME  DATE  PRCP  SNOW  SNWD  TMAX  TMIN
##     0     0  2006  8283 12746  8569  8568
```

```
oakland$DATE <- as.Date(oakland$DATE)
oakland <- oakland[order(oakland$DATE),]

oakland <- oakland[oakland$NAME %in% c('OAKLAND METROPOLITAN, CA US', 'OAKLAND MUSEUM, CA US'),]

oakland[c('SNOW', 'SNWD')][is.na(oakland[c('SNOW', 'SNWD')])] <- 0.0
# oakland <- oakland[!is.na(oakland$TMAX), ]

oakland$ID <- seq.int(nrow(oakland))
ids_to_drop <- oakland[oakland$NAME == 'OAKLAND METROPOLITAN, CA US' & oakland$DATE > '2018-06-20',]$ID
oakland <- oakland[!(oakland$ID %in% ids_to_drop), ];
oakland <- oakland[, !(colnames(oakland) == "ID")];

oakland$NAME <- 'OAKLAND'
# rearrage index column
row.names(oakland) <- NULL
```

Swarthmore

2019-09-06 - 2020-06-15

```
swarthmore_orig <- read.csv('./Data/swarthmore.csv', header = TRUE, stringsAsFactors = FALSE)

# swarthmore_orig %>% group_by(NAME) %>% summarise(n = n())
# sapply(swarthmore, function(x) sum(is.na(x)))
swarthmore <- swarthmore_orig[, keepCols_sf]

swarthmore$DATE <- as.Date(swarthmore$DATE)
swarthmore <- swarthmore[order(swarthmore$DATE),]

swarthmore <- swarthmore[swarthmore$NAME %in% c('PHILADELPHIA INTERNATIONAL AIRPORT, PA US'),]

swarthmore['SNWD'][is.na(swarthmore['SNWD'])] <- 0

swarthmore$NAME <- 'SWARTHMORE'
row.names(swarthmore) <- NULL
```

Victoria

2019-06-16 - 2020-06-15

```
victoria_orig <- read.csv('./Data/victoria.csv', header = TRUE, stringsAsFactors = FALSE)

victoria_orig %>% group_by(NAME) %>% summarise(n = n())
```

```
## # A tibble: 21 x 2
##    NAME                              n
##    <chr>                         <int>
```

```
##  1 DISCOVERY ISLAND, BC CA       2385
##  2 ESQUIMALT HARBOUR, BC CA      2420
##  3 FRIDAY HARBOR 2.6 WNW, WA US  2106
##  4 FRIDAY HARBOR 4.0 SSW, WA US  1985
##  5 FRIDAY HARBOR 4.6 WNW, WA US  2422
##  6 FRIDAY HARBOR 6.0 W, WA US    1925
##  7 FRIDAY HARBOR 6.2 WNW, WA US  2476
##  8 MALAHAT, BC CA                2479
##  9 METCHOSIN, BC CA              2220
## 10 RACE ROCKS CS, BC CA          2463
## # ... with 11 more rows
```

```
sapply(victoria_orig, function(x) sum(is.na(x)))
```

```
##   STATION      NAME  LATITUDE LONGITUDE ELEVATION      DATE      DAPR
##         0         0         0         0         0         0     36835
##      MDPR      PRCP      SNOW      SNWD      TAVG      TMAX      TMIN
##     36836      7107     22284     28797     24397     20504     22423
##      WDFG      WESD      WESF      WSFG
##     31182     37239     37226     31182
```

```
victoria <- victoria_orig[, keepCols_sf]

victoria$DATE <- as.Date(victoria$DATE)
victoria <- victoria[order(victoria$DATE),]
victoria <- victoria %>% filter(DATE >= '2019-06-16')

victoria %>% group_by(NAME) %>% summarise(prcp_na=sum(is.na(PRCP)),snow_na=sum(is.na(SNOW)), snwd_na=su
```

```
## # A tibble: 18 x 6
##    NAME                           prcp_na snow_na snwd_na tmax_na tmin_na
##    <chr>                            <int>   <int>   <int>   <int>   <int>
##  1 DISCOVERY ISLAND, BC CA            360     360     360       0       0
##  2 ESQUIMALT HARBOUR, BC CA             9     357     357       0       0
##  3 FRIDAY HARBOR 2.6 WNW, WA US         0     182     358     362     362
##  4 FRIDAY HARBOR 4.0 SSW, WA US         9     276     273     279     279
##  5 FRIDAY HARBOR 4.6 WNW, WA US         1     168     352     352     352
##  6 FRIDAY HARBOR 6.0 W, WA US          29     136     275     276     276
##  7 FRIDAY HARBOR 6.2 WNW, WA US         0     174     358     360     360
##  8 MALAHAT, BC CA                     331     360     360       0       0
##  9 METCHOSIN, BC CA                    15      15      19     253     253
## 10 RACE ROCKS CS, BC CA               355     355     355       0       0
## 11 SAANICHTON CDA, BC CA               19      19      18      19      20
## 12 SAANICHTON MOUNT NEWTON, BC CA      20      20      37     261     261
## 13 VICTORIA 0.9 E, CA                  38     128     128     128     128
## 14 VICTORIA 2.9 W, CA                  14     139     316     316     316
## 15 VICTORIA 5.9 N, CA                   2       4      59     350     350
## 16 VICTORIA 6.0 NNW, CA                 1       1      43     361     361
## 17 VICTORIA GONZALES HTS, BC CA         4     362     362     172     172
## 18 VICTORIA UNIVERSITY CS, BC CA       29     361     361       0       0
```

```
victoria <- victoria %>% filter(NAME %in% c('VICTORIA UNIVERSITY CS, BC CA', 'VICTORIA 6.0 NNW, CA'))
victoria$NAME <- 'VICTORIA'

victoria <- aggregate(victoria, by=list(DATE_ID=victoria$DATE), min, na.rm = TRUE)

victoria[is.infinite(victoria$SNWD),]['SNWD'] <- 0
victoria[is.infinite(victoria$SNOW),]['SNOW'] <- 0
sapply(victoria, function(x) sum(is.na(x)))
```

```
## DATE_ID    NAME    DATE    PRCP    SNOW    SNWD    TMAX    TMIN
##       0       0       0       0       0       0       0       0
```

```
victoria <- victoria[, !(colnames(victoria)=='DATE_ID')]
```

Leuven

2019-06-15 - 2020-06-14

Notes:
- Data for Swarthmore are from Philadelphia International Airport Station, which is closest to Swarthmore.