Guide 2: Cleaning and Pre-Processing Weather Station Data

Marc Los Huertos

February 2, 2024 (ver. 0.4)

1 Introduction

1.1 Goals

The goal of this guide is to provide a step-by-step process for cleaning and pre-processing weather station data. The data is from the Global Historical Climatology Network (GHCN) Daily dataset. The data is available from the National Oceanic and Atmospheric Administration (NOAA) and is available from the National Centers for Environmental Information (NCEI) at

1.2 Background

What is GHCN-Daily?Links to an external site.

1.3 Approach

2 Cleaning and Pre-Processing Functions

2.1 Starting the Process

Before you begin, make sure you have the stations read into R. You can do this by running the following code:

```
#ls()
```

You should see station1...stationN, where N is the number of stations you have read in.

Using str(), make sure the data sets look right!

```
## Error in source(here("04_Regional_Climate_Trends", "Guides", "Guide2.R")):
/home/mwl04747/RTricks/04_Regional_Climate_Trends/Guides/Guide2.R:17:0:
unexpected end of input
## 15:
```

```
## 16:
##    ^
```

2.2 Clean Data

First, I tested each line on station1. I will then create a function to clean the data and apply it to each station.

```
# station1£VALUE = station1£VALUE/10  # Correct Values Units
# Fix Date format
station1$Ymd = as.Date(as.character(station1$DATE), format = "%Y%m%d")

## Error in as.Date(as.character(station1$DATE), format = "%Y%m%d"):
object 'station1' not found

str(station1)

## Error in str(station1):  object 'station1' not found

station1$MONTH = as.numeric(format(station1$Ymd, "%m"))

## Error in format(station1$Ymd, "%m"):  object 'station1' not found

station1$YEAR = as.numeric(format(station1$Ymd, "%Y"))

## Error in format(station1$Ymd, "%Y"):  object 'station1' not found

station1.monthly = aggregate(VALUE ~ MONTH + YEAR,
                   data = subset(station1, ELEMENT == "TMAX"), mean)

## Error in subset(station1, ELEMENT == "TMAX"):  object 'station1'
not found

# create baseline (normals) dataset
station1.normals = subset(station1,
                          Ymd >= "1961-01-01" & Ymd <= "1990-12-31")

## Error in subset(station1, Ymd >= "1961-01-01" & Ymd <= "1990-12-31"):
object 'station1' not found

station1.normals.monthly = aggregate(VALUE ~ MONTH,
                   data = subset(station1.normals, ELEMENT == "TMAX"), mean)

## Error in subset(station1.normals, ELEMENT == "TMAX"):  object 'station1.normals'
not found

names(station1.normals.monthly) <- c("MONTH", "NORMALS")

## Error in names(station1.normals.monthly) <- c("MONTH", "NORMALS"):
object 'station1.normals.monthly' not found
```

```
station1.anomaly = merge(station1.monthly,
                         station1.normals.monthly, by = "MONTH")

## Error in merge(station1.monthly, station1.normals.monthly, by =
"MONTH"):  object 'station1.monthly' not found

station1.anomaly$ANOMALY =
  station1.anomaly$VALUE - station1.anomaly$NORMALS

## Error in eval(expr, envir, enclos):  object 'station1.anomaly' not
found
```

2.3 Clean Data Function

Function is probably sensitive to missing values, need to work on that!

```
x=station1

## Error in eval(expr, envir, enclos):  object 'station1' not found

cleandataframe.fun <- function(x){
  #x£VALUE = x£VALUE/10
  x$Ymd = as.Date(as.character(x$DATE), format = "%Y%m%d")
  x$MONTH = as.numeric(format(x$Ymd, "%m"))
  x$YEAR = as.numeric(format(x$Ymd, "%Y"))

  x.TMAX.monthly = aggregate(VALUE ~ MONTH + YEAR,
            data = subset(x, ELEMENT == "TMAX"), mean)
  names(x.TMAX.monthly) <- c("MONTH", "YEAR", "TMAX")
  x.TMIN.monthly = aggregate(VALUE ~ MONTH + YEAR,
            data = subset(x, ELEMENT == "TMIN"), mean)
  names(x.TMIN.monthly) <- c("MONTH", "YEAR", "TMIN")
  x.PRCP.monthly = aggregate(VALUE ~ MONTH + YEAR,
            data = subset(x, ELEMENT == "PRCP"), sum)
  names(x.PRCP.monthly) <- c("MONTH", "YEAR", "PRCP")

  x.normals = subset(x, Ymd >= "1961-01-01" & Ymd <= "1990-12-31")
  x.TMAX.normals.monthly = aggregate(VALUE ~ MONTH,
            data = subset(x.normals, ELEMENT == "TMAX"), mean)
  names(x.TMAX.normals.monthly) <- c("MONTH", "NORMALS")
  x.TMIN.normals.monthly = aggregate(VALUE ~ MONTH,
            data = subset(x.normals, ELEMENT == "TMIN"), mean)
  names(x.TMIN.normals.monthly) <- c("MONTH", "NORMALS")
  x.PRCP.normals.monthly = aggregate(VALUE ~ MONTH,
            data = subset(x.normals, ELEMENT == "PRCP"), sum)
  names(x.PRCP.normals.monthly) <- c("MONTH", "NORMALS")
```

```
  x.TMAX.anomaly = merge(x.TMAX.monthly, x.TMAX.normals.monthly, by = "MONTH")
  x.TMAX.anomaly$TMAX.anomaly = x.TMAX.anomaly$TMAX - x.TMAX.anomaly$NORMALS

  x.TMIN.anomaly = merge(x.TMIN.monthly, x.TMIN.normals.monthly, by = "MONTH")
  x.TMIN.anomaly$TMIN.anomaly = x.TMIN.anomaly$TMIN - x.TMIN.anomaly$NORMALS

  x.PRCP.anomaly = merge(x.PRCP.monthly, x.PRCP.normals.monthly, by = "MONTH")
  x.PRCP.anomaly$PRCP.anomaly = x.PRCP.anomaly$PRCP - x.PRCP.anomaly$NORMALS

  TEMP <- merge(x.TMAX.anomaly, x.TMIN.anomaly, by = c("MONTH", "YEAR") )
  x.anomaly <- merge(TEMP, x.PRCP.anomaly, by = c("MONTH", "YEAR"))[,c(1:3, 5:6, 8:9, 11)]
  library(lubridate)
  #x.anomaly£Ym1 = as.Date(paste(x.anomaly£YEAR, x.anomaly£MONTH), format="%Y %m")
  x.anomaly$Ym1 =  lubridate::myd(paste(x.anomaly$MONTH, x.anomaly$YEAR, "1"))
  str(x.anomaly)
  return(x.anomaly)
}
```

2.4 Apply Function to All Stations

So far, I have only run function for 1 station, but I suspect you can figure out how to run it for each one!

```
station1.clean= cleandataframe.fun(station1)

## Error in as.Date(as.character(x$DATE), format = "%Y%m%d"):  object
'station1' not found
```

2.5 Plot Anomaly

Graphic has lots of issues. more next time! But here's a start.

```
options(scipen=5)
par(mar=c(4,6,2,5))

plot(ANOMALY ~ YEAR, data = subset(station1.TMAX, MONTH == 1),
     las=1, pch=19, col = "blue", cex=.5, #xlab = "Year",
     ylab = "Maximum Temp Anomaly (C)",
     main="January Maximum Temp Anomaly")

## Error in subset(station1.TMAX, MONTH == 1):  object 'station1.TMAX'
not found
```

```
mtext("Maximum Temp Anomaly (C)", side = 2, line = 3)

## Error in mtext("Maximum Temp Anomaly (C)", side = 2, line = 3):
plot.new has not been called yet

temp.lm = lm(ANOMALY ~ YEAR, data = subset(station1.TMAX, MONTH == 1))

## Error in subset(station1.TMAX, MONTH == 1):  object 'station1.TMAX'
not found

abline(coef(temp.lm), col = "red")

## Error in coef(temp.lm):  object 'temp.lm' not found
```

3 QA/QC

3.1 Missing Data

```
# determine percent missing in station1
station1.TMAX.coverage = sum(!is.na(station1$VALUE[station1$ELEMENT=="TMAX"]))/length(static

## Error in eval(expr, envir, enclos):  object 'station1' not found

# function to determine percent missing
coverage.fun <- function(station, element){
  Dates.all = data.frame(Ymd=seq.Date(from=min(station$Ymd), to=max(station$Ymd), by="day"))
  station.full = merge(Dates.all, station, all = TRUE)
  station.coverage = sum(!is.na(station.full$VALUE[station.full$ELEMENT==element]))/
    length(station.full$VALUE[station.full$ELEMENT==element])*100
  return(round(station.coverage,2))
}

coverage.data(station1, "TMAX")

## Error in coverage.data(station1, "TMAX"):  could not find function
"coverage.data"

coverage.data(station2, "TMAX")

## Error in coverage.data(station2, "TMAX"):  could not find function
"coverage.data"

Date.full = data.frame(Ymd=seq.Date(from=min(station1$Ymd), to=max(station1$Ymd), by="day"))

## Error in seq.Date(from = min(station1$Ymd), to = max(station1$Ymd),
by = "day"):  object 'station1' not found

str(Date.full)
```

```
## Error in str(Date.full):  object 'Date.full' not found
```

```
station1.full = merge(Date.full, station1, all = TRUE)
```

```
## Error in merge(Date.full, station1, all = TRUE): object 'Date.full'
not found
```

```
coverage.fun(station1, "TMAX")
```

```
## Error in seq.Date(from = min(station$Ymd), to = max(station$Ymd),
by = "day"):  object 'station1' not found
```

```
coverage.fun(station2, "TMAX")
```

```
## Error in seq.Date(from = min(station$Ymd), to = max(station$Ymd),
by = "day"):  object 'station2' not found
```

4 Next Steps

This is all we need to do so far. Next week, we'll look at different way to visualize the data!

I'll save all the station data into csv files, then use them in the next guide to clean, process, and visualize data.

```
write.csv(station1, file = paste0(here::here("04_Regional_Climate_Trends", "Data", "SP24", '
```

```
## Error in is.data.frame(x):  object 'station1' not found
```

```
write.csv(station2, file = paste0(here::here("04_Regional_Climate_Trends", "Data", "SP24", '
```

```
## Error in is.data.frame(x):  object 'station2' not found
```

```
write.csv(station3, file = paste0(here::here("04_Regional_Climate_Trends", "Data", "SP24", '
```

```
## Error in is.data.frame(x):  object 'station3' not found
```

```
write.csv(station4, file = paste0(here::here("04_Regional_Climate_Trends", "Data", "SP24", '
```

```
## Error in is.data.frame(x):  object 'station4' not found
```

```
write.csv(station5, file = paste0(here::here("04_Regional_Climate_Trends", "Data", "SP24", '
```

```
## Error in is.data.frame(x):  object 'station5' not found
```