Basic Data Management

Convert character to date-time

https://github.com/bozhang22/CARE/tree/main/SampleData

- tidyverse
- lubridate

```
Edgewater_B_2021$Time <- ymd_hms(Edgewater_B_2021$created_at, tz = "UTC")
plot(Edgewater_B_2021$Time , Edgewater_B_2021$PM2.5_CF1_ug.m3, type = "l",
ylim = c(0, 400))
Edgewater$Time <- ymd_hms(Edgewater$created_at, tz = "UTC")

Edgewater_B_2021$Time_local <- with_tz(Edgewater_B_2021$Time,
"America/Chicago")
Edgewater$Time_local <- with_tz(Edgewater$Time, "America/Chicago")

plot(Edgewater$Time_local, Edgewater$PM2.5_CF1_ug.m3)
```

Recoding variables

- variable[condition] <- expression
- E.g., use 50 as a threshold

```
Edgewater_B_20210101$level_2.5[Edgewater_B_20210101$PM2.5_CF1_ug.m3 > 50] <- "High"
Edgewater_B_20210101$level_2.5[Edgewater_B_20210101$PM2.5_CF1_ug.m3 <= 50] <- "Low"

Edgewater_B_20210101$level_2.5 <- as.factor(Edgewater_B_20210101$level_2.5)
summary(Edgewater_B_20210101$level_2.5)
```

Rename variables

- names()

```
names(Edgewater)
names(Edgewater)[2] <- "PM1.0"
names(Edgewater)[2:4] <- c("PM1.0", "PM2.5", "PM10")
names(Edgewater)
```

Missing values

- In R, missing values are represented by the symbol NA (not available).
- Some functions don't work with NA value.

```
mean(Edgewater_B_2021$Pressure_hpa)
mean(Edgewater_B_2021$Pressure_hpa, na.rm = T)
```

Subsetting datasets

- Subsetting based on columns and rows

```
myvars <- c("Time", "PM1.0", "PM2.5", "PM10")
newdata <- Edgewater[myvars]
Edgewater$IAQ <- NULL

startdate <- ymd_hms("2021-02-01 00:00:00", tz = "UTC")
enddate <- ymd_hms("2021-02-28 23:59:59", tz = "UTC")
Feb_data <- Edgewater_B_2021[which(Edgewater_B_2021$Time >= startdate &
Edgewater_B_2021$Time <= enddate),]
pm_8 <- Edgewater_B_2021[which(hour(Edgewater_B_2021$Time) == 20),]
```

Practice

Use the data you downloaded to finish the following tasks:

- Create a new variable which stores the local time of the record.
- Create a time series plot with one line representing pm2.5 and another line representing pm10.
- Recode pm10 to “high” and “low” using a threshold of 40. How many records have a value of “high” and how many have a value of “low”?
- Calculate the mean pm2.5 and pm10 level based on all the records in your data frame.
- Calculate the mean pm2.5 and pm10 level based on records collected between 7am-9am in your data frame.