# データ収集

東京国際大学 データサイエンス教育研究所 竹田 恒

2022-08-28



# 1 取得するデータ

気象庁の過去の気象データ検索 (https://www.data.jma.go.jp/obd/stats/etrn/index.php?prec_no=44&block_no=47662&year=2022&month=08&day=10&view=)から東京の2022年8月10日の気温と風向を取得す

る.
東京 2022年8月10日（1時間ごとの値） (https://www.data.jma.go.jp/obd/stats/etrn/view/hourly_s1.php?prec_no=44&block_no=47662&year=2022&month=8&day=10&view=)

# 1時間ごとの値



東京 2022年8月10日 （1時間ごとの値）

| 時 | 気圧(hPa) | | 降水量 | 気温 | 露点温度 | 蒸気圧 | 湿度 | 風向・風速(m/s) | | 日照時間 | 全天日射量 | 雪(cm) | | 天気 | 雲量 | 視程 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 現地 | 海面 | (mm) | (℃) | (℃) | (hPa) | (%) | 風速 | 風向 | (h) | (MJ/m²) | 降雪 | 積雪 | | | (km) |
| 1 | 1007.0 | 1009.7 | -- | 28.7 | 23.2 | 28.4 | 72 | 3.8 | 南 | | | × | × | | | |
| 2 | 1006.5 | 1009.2 | -- | 28.7 | 23.2 | 28.4 | 72 | 4.8 | 南 | | | × | × | | | |
| 3 | 1006.9 | 1009.6 | -- | 28.5 | 23.2 | 28.4 | 73 | 5.2 | 南 | | | × | × | ◐ | 2 | 20.0 |
| 4 | 1006.7 | 1009.4 | -- | 28.2 | 23.1 | 28.3 | 74 | 4.0 | 南 | | | × | × | | | |
| 5 | 1007.0 | 1009.7 | -- | 27.8 | 23.4 | 28.8 | 77 | 4.1 | 南 | 0.0 | 0.00 | × | × | | | |
| 6 | 1007.3 | 1010.0 | -- | 28.0 | 23.6 | 29.1 | 77 | 3.9 | 南 | 0.3 | 0.17 | × | × | ◐ | 5 | 20.0 |
| 7 | 1007.4 | 1010.1 | -- | 29.6 | 23.6 | 29.0 | 70 | 3.4 | 南 | 1.0 | 0.89 | × | × | | | |
| 8 | 1007.6 | 1010.3 | -- | 31.2 | 23.8 | 29.6 | 65 | 4.9 | 南南西 | 1.0 | 1.66 | × | × | | | |
| 9 | 1007.5 | 1010.2 | -- | 32.5 | 23.5 | 28.9 | 59 | 5.1 | 南南西 | 1.0 | 2.29 | × | × | ◐ | 5 | 25.0 |
| 10 | 1007.5 | 1010.2 | -- | 33.4 | 23.7 | 29.3 | 57 | 4.4 | 南南西 | 1.0 | 2.84 | × | × | | | |
| 11 | 1007.2 | 1009.9 | -- | 34.3 | 24.3 | 30.3 | 56 | 6.5 | 南南西 | 1.0 | 3.26 | × | × | | | |
| 12 | 1006.9 | 1009.6 | -- | 34.3 | 24.3 | 30.3 | 56 | 5.4 | 南南西 | 1.0 | 3.26 | × | × | ◐ | 5 | 25.0 |
| 13 | 1006.6 | 1009.3 | -- | 34.3 | 25.1 | 31.9 | 59 | 6.9 | 南 | 1.0 | 2.81 | × | × | | | |
| 14 | 1006.5 | 1009.2 | -- | 33.8 | 25.8 | 33.2 | 63 | 5.7 | 南 | 1.0 | 2.81 | × | × | | | |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 1006.4 | 1009.1 | -- | 33.6 | 25.0 | 31.8 | 61 | 7.1 | 南 | 1.0 | 2.57 | × | × | ◐ | 7 | 20.0 |
| 16 | 1006.5 | 1009.2 | -- | 32.6 | 25.4 | 32.5 | 66 | 4.3 | 南 | 1.0 | 1.82 | × | × | | | |
| 17 | 1006.6 | 1009.3 | -- | 31.3 | 25.2 | 32.0 | 70 | 5.1 | 南南西 | 1.0 | 1.16 | × | × | | | |
| 18 | 1007.0 | 1009.7 | -- | 30.2 | 25.1 | 31.8 | 74 | 6.9 | 南 | 0.8 | 0.41 | × | × | ● | 10- | 20.0 |
| 19 | 1007.2 | 1009.9 | -- | 29.6 | 24.9 | 31.5 | 76 | 4.1 | 南 | 0.0 | 0.04 | × | × | | | |
| 20 | 1008.0 | 1010.7 | -- | 28.7 | 25.3 | 32.3 | 82 | 5.0 | 南南東 | | 0.00 | × | × | | | |
| 21 | 1008.3 | 1011.0 | -- | 28.5 | 25.1 | 31.9 | 82 | 4.3 | 南 | | | × | × | ● | 10- | 20.0 |
| 22 | 1008.1 | 1010.8 | -- | 28.7 | 24.7 | 31.1 | 79 | 5.3 | 南 | | | × | × | | | |
| 23 | 1008.3 | 1011.0 | -- | 28.2 | 25.0 | 31.8 | 83 | 5.8 | 南 | | | × | × | | | |
| 24 | 1008.4 | 1011.1 | -- | 28.1 | 24.9 | 31.6 | 83 | 4.4 | 南 | | | × | × | | | |

出典：気象庁｜過去の気象データ検索

# 2 URL解析

取得したいデータがあるウェブページのURLを見てどのような規則で表示されているかを推察する．

```
https://www.data.jma.go.jp/obd/stats/etrn/view/hourly_s1.php?prec_no=44&block_no=47662&year=2022&month=8&day=10&view=
```

この例では，prec_no，block_noは恐らく気象観測所に関する番号であろう，year，month，dayが年月日だろうと当たりが付く．このようにパラメータを含む形でURLが記載されている場合でページソースにHTMLテーブルやXMLの形式でデータがある場合（ウェブ画面を右クリックしソースを表示すると分かる）は，パラメータをプログラムで変えることで異なるウェブページに機械的にアクセスしデータを取得できる．ただし，Yahoo！ファイナンスのようにウェブサーバーの負荷の問題からウェブスクレイピングを禁止してるところもあるので利用規約を読むことが必要．ウェブスクレイピング自体は合法的である．

cf. スクレイピングは違法？Webスクレイピングに関する10のよくある誤解！ (https://www.octoparse.jp/blog/10-myths-about-web-scraping)

# 3 設定

```
# 出力ファイル
DB  <- 'weather.db'  # データベース名
F.O <- 'weather.csv' # CSVファイル名

# 気象観測所
site <- data.frame(
  id   = 47662,   # 番号
  name = 'Tokyo') # 名称（データベースのテーブル名として使う）

# システムロケール（海外クラウド環境利用時に必要な時間と言語の設定）
Sys.setlocale("LC_TIME", "ja_JP.UTF-8")
```

```
## [1] "ja_JP.UTF-8"
```

```
# 対象日時（テーブル取得のためのURLに適用する日時）
lt <- as.POSIXlt('2022-08-10') # POSIX準拠ローカル時間
year  <- 1900 + lt$year
month <- 1 + lt$mon
day   <- lt$mday
```

# 4 URL作成

```
url <- paste0('https://www.data.jma.go.jp/obd/stats/etrn/view/hourly_s1.php?prec_no=44&bloc
k_no=', site$id, '&year=', year, '&month=', month, '&day=', day, '&view=')

cat('URL:', url, fill = T) # 作成したURLを表示
```

```
## URL:
## https://www.data.jma.go.jp/obd/stats/etrn/view/hourly_s1.php?prec_no=44&block_no=47662&
year=2022&month=8&day=10&view=
```

# 5 ウェブページのデータ取得 (Web scraping)

# 5.1 テーブル取得

```
library(rvest)
tbl <- read_html(url) %>% html_table()
tbl
```

```
## [[1]]
## # A tibble: 1 × 2
##   X1    X2
##   <lgl> <lgl>
## 1 NA    NA
##
## [[2]]
## # A tibble: 1 × 2
##   X1    X2
##   <lgl> <lgl>
## 1 NA    NA
##
## [[3]]
## # A tibble: 1 × 7
##   X1    X2    X3    X4    X5    X6    X7
##   <lgl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl>
## 1 NA    NA    NA    NA    NA    NA    NA
##
## [[4]]
## # A tibble: 1 × 3
##   X1    X2    X3
##   <lgl> <lgl> <lgl>
## 1 NA    NA    NA
##
```

```
## [[5]]
## # A tibble: 25 × 17
##    時    気圧(…¹ 気圧(…² 降水…³  気温(…⁴ 露点…⁵  蒸気…⁶  湿度(…⁷ 風向…⁸ 風向・…⁹
##    <chr> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>  <chr>
##  1 時    現地    海面    降水量… 気温(℃) 露点温… 蒸気圧… 湿度(… 風速   風向
##  2 1     1007.0  1009.7  --      28.7    23.2    28.4    72      3.8    南
##  3 2     1006.5  1009.2  --      28.7    23.2    28.4    72      4.8    南
##  4 3     1006.9  1009.6  --      28.5    23.2    28.4    73      5.2    南
##  5 4     1006.7  1009.4  --      28.2    23.1    28.3    74      4.0    南
##  6 5     1007.0  1009.7  --      27.8    23.4    28.8    77      4.1    南
##  7 6     1007.3  1010.0  --      28.0    23.6    29.1    77      3.9    南
##  8 7     1007.4  1010.1  --      29.6    23.6    29.0    70      3.4    南
##  9 8     1007.6  1010.3  --      31.2    23.8    29.6    65      4.9    南南西
## 10 9     1007.5  1010.2  --      32.5    23.5    28.9    59      5.1    南南西
## # … with 15 more rows, 7 more variables: `日照時間(h)` <chr>,
## #   `全天日射量(MJ/㎡)` <chr>, `雪(cm)` <chr>, `雪(cm)` <chr>, 天気 <chr>,
## #   雲量 <chr>, `視程(km)` <chr>, and abbreviated variable names ¹`気圧(hPa)`,
## #   ²`気圧(hPa)`, ³`降水量(mm)`, ⁴`気温(℃)`, ⁵`露点温度(℃)`, ⁶`蒸気圧(hPa)`,
## #   ⁷`湿度(%)`, ⁸`風向・風速(m/s)`, ⁹`風向・風速(m/s)`
## # ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
##
## [[6]]
## # A tibble: 1 × 1
##   X1
```

```
##   <chr>
## 1 ーー
##
## [[7]]
## # A tibble: 1 × 4
##   X1             X2                  X3                  X4
##   <chr>          <chr>               <chr>               <chr>
## 1 利用される方へ よくある質問（FAQ） 気象観測統計の解説 年・季節・各月の天候
```

tblをプリントし，どのテーブルに必要なデータが格納されているか確認すると5番目のテーブル（リスト）にあることが分かる．次で5番目のリストを取り出す．すべてテキストデータ(chr)になっている．

```
d0 <- as.data.frame(tbl[[5]])
str(d0)
```

```
## 'data.frame':    25 obs. of  17 variables:
##  $ 時                : chr  "時" "1" "2" "3" ...
##  $ 気圧(hPa)         : chr  "現地" "1007.0" "1006.5" "1006.9" ...
##  $ 気圧(hPa)         : chr  "海面" "1009.7" "1009.2" "1009.6" ...
##  $ 降水量(mm)        : chr  "降水量(mm)" "--" "--" "--" ...
##  $ 気温(℃)           : chr  "気温(℃)" "28.7" "28.7" "28.5" ...
##  $ 露点温度(℃)       : chr  "露点温度(℃)" "23.2" "23.2" "23.2" ...
##  $ 蒸気圧(hPa)       : chr  "蒸気圧(hPa)" "28.4" "28.4" "28.4" ...
##  $ 湿度(%)           : chr  "湿度(%)" "72" "72" "73" ...
##  $ 風向・風速(m/s)   : chr  "風速" "3.8" "4.8" "5.2" ...
##  $ 風向・風速(m/s)   : chr  "風向" "南" "南" "南" ...
##  $ 日照時間(h)       : chr  "日照時間(h)" "" "" "" ...
##  $ 全天日射量(MJ/㎡): chr  "全天日射量(MJ/㎡)" "" "" "" ...
##  $ 雪(cm)            : chr  "降雪" "×" "×" "×" ...
##  $ 雪(cm)            : chr  "積雪" "×" "×" "×" ...
##  $ 天気              : chr  "天気" "" "" "" ...
##  $ 雲量              : chr  "雲量" "" "" "2" ...
##  $ 視程(km)          : chr  "視程(km)" "" "" "20.0" ...
```

# 5.2 テーブルの整形

日時などの情報追加，変数の型指定，データの整形を行い，書込用テーブルを作成する．日時のフォーマット（%Y-%m-%d %H:%M:%Sなど)はどのプログラミング言語でもほぼ共通なのでしっかりと記憶して

おくこと．Rコンソールで「?strptime」とタイプすれば他の記号も調べることができる．

```
# 日時整形
hour <- d0[-1, '時'] # 1列目は時刻1〜24（-1:一行目は不要なため削除）
# コンピュータの世界(POSIX準拠)では24時は存在しないので0〜23にする必要がある.
# コンピュータ上では24時は翌日の日付になる.
datetime <- as.POSIXlt(paste(lt, hour),        # 例）2022-08-10 24
                       format = '%Y-%m-%d %H') # 例の様な数字を「年-月-日 時」として読込む
                                               # 自動で時刻が0〜23に変換される.

# 書込用テーブル作成
d1 <- data.frame(site.id   = as.integer(site$id), # 整数型
                 site.name = site$name,
                 datetime  = format(datetime, '%Y-%m-%d %H:00'),
                 temp      = as.double(d0[-1, 5]), # 倍精度浮動小数点型
                 wind      = d0[-1, 10])
str(d1)
```

```
## 'data.frame':    24 obs. of  5 variables:
##  $ site.id  : int  47662 47662 47662 47662 47662 47662 47662 47662 47662 47662 ...
##  $ site.name: chr  "Tokyo" "Tokyo" "Tokyo" "Tokyo" ...
##  $ datetime : chr  "2022-08-10 01:00" "2022-08-10 02:00" "2022-08-10 03:00" "2022-08-10
04:00" ...
##  $ temp     : num  28.7 28.7 28.5 28.2 27.8 28 29.6 31.2 32.5 33.4 ...
##  $ wind     : chr  "南" "南" "南" "南" ...
```

# 6 データ保存

## 6.1 データベース（SQLite）への保存

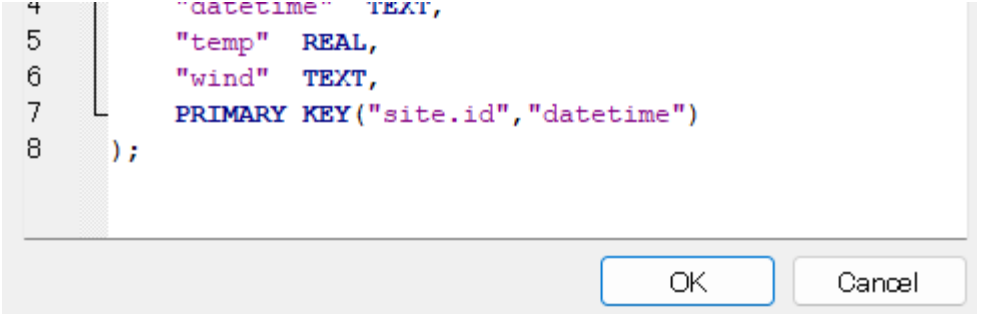
```
library(RSQLite)

# データベース接続
conn <- dbConnect(RSQLite::SQLite(), DB)

# 既存テーブル削除（必要に応じて実施）
dbSendQuery(conn, paste('DROP TABLE IF EXISTS', site$name))
```

```
## <SQLiteResult>
##   SQL  DROP TABLE IF EXISTS Tokyo
##   ROWS Fetched: 0 [complete]
##        Changed: 0
```

```
# テーブル追記書込
dbWriteTable(conn, site$name, d1, append = T)
```

レコードを重複して保存したくない場合は，レコードがユニークになるように PRIMARY KEYを設定するとよい．DB Browser for SQLiteなどのGUIツールを使うと簡単に設定できる．設定すると重複するレコードが来たときにプログラムはエラー停止する．繰返計算などで停止させたくないときはtry()関数を付けるとプログラム実行を継続できる．


テーブルの定義を編集
?
テーブル
Tokyo
高度な設定
フィールド
制約
追加
削除
先頭へ
上へ
下へ
末尾へ
名前
データ型
NN
PK
AI
U
site.id
INTEGER
site.name
TEXT
datetime
TEXT
temp
REAL
wind
TEXT
1 CREATE TABLE "Tokyo" (
2 "site.id" INTEGER,
3 "site.name" TEXT,




PRIMARY KEYの設定例（DB Browser for SQLite）

複数のカラムを組み合わせたPRIMARY KEY（PK）とすることでレコードがユニークになる．また，データハンドリングも高速になる利点もある．

```
# データ選択（ちゃんと保存されたか確認すること）
res <- dbSendQuery(conn, 'SELECT * FROM Tokyo')

# 選択結果取得
dbFetch(res)
```

```
##    site.id site.name         datetime temp   wind
## 1    47662     Tokyo 2022-08-10 01:00 28.7     南
## 2    47662     Tokyo 2022-08-10 02:00 28.7     南
## 3    47662     Tokyo 2022-08-10 03:00 28.5     南
## 4    47662     Tokyo 2022-08-10 04:00 28.2     南
## 5    47662     Tokyo 2022-08-10 05:00 27.8     南
## 6    47662     Tokyo 2022-08-10 06:00 28.0     南
## 7    47662     Tokyo 2022-08-10 07:00 29.6     南
## 8    47662     Tokyo 2022-08-10 08:00 31.2 南南西
## 9    47662     Tokyo 2022-08-10 09:00 32.5 南南西
## 10   47662     Tokyo 2022-08-10 10:00 33.4 南南西
## 11   47662     Tokyo 2022-08-10 11:00 34.3 南南西
## 12   47662     Tokyo 2022-08-10 12:00 34.3 南南西
## 13   47662     Tokyo 2022-08-10 13:00 34.3     南
## 14   47662     Tokyo 2022-08-10 14:00 33.8     南
## 15   47662     Tokyo 2022-08-10 15:00 33.6     南
## 16   47662     Tokyo 2022-08-10 16:00 32.6     南
## 17   47662     Tokyo 2022-08-10 17:00 31.3 南南西
## 18   47662     Tokyo 2022-08-10 18:00 30.2     南
## 19   47662     Tokyo 2022-08-10 19:00 29.6     南
## 20   47662     Tokyo 2022-08-10 20:00 28.7 南南東
## 21   47662     Tokyo 2022-08-10 21:00 28.5     南
## 22   47662     Tokyo 2022-08-10 22:00 28.7     南
## 23   47662     Tokyo 2022-08-10 23:00 28.2     南
```

```
## 24   47662     Tokyo 2022-08-11 00:00 28.1     南
```

```
# 選択結果解放
dbClearResult(res)

# データベース接続解除
dbDisconnect(conn)
```

# 6.2 CSVファイルへの保存

```
# 既存ファイル削除（必要に応じて実施）
file.remove(F.O)
```

```
## [1] TRUE
```

```
# テーブル追記書込
library(data.table)
fwrite(d1, file = F.O, sep = ',', append = T)
```

```
# 読込確認
(d2 <- fread(file = F.O))
```

```
##     site.id site.name         datetime temp   wind
##  1:   47662     Tokyo 2022-08-10 01:00 28.7     南
##  2:   47662     Tokyo 2022-08-10 02:00 28.7     南
##  3:   47662     Tokyo 2022-08-10 03:00 28.5     南
##  4:   47662     Tokyo 2022-08-10 04:00 28.2     南
##  5:   47662     Tokyo 2022-08-10 05:00 27.8     南
##  6:   47662     Tokyo 2022-08-10 06:00 28.0     南
##  7:   47662     Tokyo 2022-08-10 07:00 29.6     南
##  8:   47662     Tokyo 2022-08-10 08:00 31.2 南南西
##  9:   47662     Tokyo 2022-08-10 09:00 32.5 南南西
## 10:   47662     Tokyo 2022-08-10 10:00 33.4 南南西
## 11:   47662     Tokyo 2022-08-10 11:00 34.3 南南西
## 12:   47662     Tokyo 2022-08-10 12:00 34.3 南南西
## 13:   47662     Tokyo 2022-08-10 13:00 34.3     南
## 14:   47662     Tokyo 2022-08-10 14:00 33.8     南
## 15:   47662     Tokyo 2022-08-10 15:00 33.6     南
## 16:   47662     Tokyo 2022-08-10 16:00 32.6     南
## 17:   47662     Tokyo 2022-08-10 17:00 31.3 南南西
## 18:   47662     Tokyo 2022-08-10 18:00 30.2     南
## 19:   47662     Tokyo 2022-08-10 19:00 29.6     南
## 20:   47662     Tokyo 2022-08-10 20:00 28.7 南南東
## 21:   47662     Tokyo 2022-08-10 21:00 28.5     南
## 22:   47662     Tokyo 2022-08-10 22:00 28.7     南
## 23:   47662     Tokyo 2022-08-10 23:00 28.2     南
```

```
## 24:   47662     Tokyo 2022-08-11 00:00 28.1     南
##     site.id site.name         datetime temp   wind
```

# 7 演習課題

次をおこなうRソースファイル（拡張子：R）を作成，プログラム実行しデータベースを作成せよ．また，値が無いところはどうすれば良いか検討せよ．

> 2021年12月31日～2022年1月1日までの気象データ（気温，湿度，日照時間，風向）をデータベースに格納する．

ただし，forループを使いデータを取得すること．

【重要】プログラムで連続してデータ収集する場合は，ウェブサーバーの負荷軽減のため，プログラムを休止させるコマンド：Sys.sleep(runif(1, min = 1, max = 2))をループ内に置き1～2秒の間隔でデータを取得すること．（忘れたら0点）

ヒント：

```
lt.fr <- as.POSIXlt('2021-12-30')
lt.to <- as.POSIXlt('2022-01-01')
lst   <- as.POSIXlt(seq(lt.fr, lt.to, by = 'days'))
lst
```

```
## [1] "2021-12-30 JST" "2021-12-31 JST" "2022-01-01 JST"
```