DATA.TABLE

USING `data.table`

adv: quick responsive efficient for very large data sets
-other universe (other than tidyverse) - adaptations for R when working with large datasets

INSTALL AND LOAD PACKAGES

```
pacman::p_load(pacman, data.table, httr, tidyverse)
```

- **data.table** - for working with large datasets
- **httr** - for working with web sources (need live data connection to do this)

LOAD AND PREPARE DATA

```
browseURL("https://figshare.com")
```

- site for sharing scientific data
- weather data in csv format (43 MB) N = 655,362

Set url where data is

```
url <- "https://ndownloader.figshare.com/files/8448380"
```

- results:

environment - values: url "https://ndownloader

Values	
url	"https://ndownloader.figshare.com/files/8448380"

Read URL (US daily weather data by station)

```
response <- httr::HEAD(url)
```

- results:

environment - data: response list of 10
- click on list > tells us what types of data

Data	
response	List of 10

Get the file size in bytes and confirm

```
httr::headers(response)[[Content-Length"]]
```

- results:

console: 43 mb

Download with ***`data.table`***

```
df <- fread(url)
```

- results:

can take a bit - environ - data: df 655362 obs of 5 variables

Data	
df	655362 obs. of 5 variables
response	List of 10

Take a quick look at the data

```
head(df)
```

- results:

1: V1 (index number) / data / date / param / siteid

```
> # Take a look at the data
> head(df)
   V1  data       date          param    siteid
1:  1  0.00 2003-01-01  Precipitation      ACRE
2:  2  0.00 2003-01-02  Precipitation AlbertLea
3:  3 11.32 2003-01-03  Precipitation      Ames
4:  4  0.00 2003-01-04  Precipitation    Antigo
5:  5  3.04 2003-01-05  Precipitation  Appleton
6:  6  0.49 2003-01-06  Precipitation Arlington
```

WORK IN DATA.TABLE

data.table notation> **`df[i, j, by]`**

- **`i`**: filter by row (empty of no filter is needed)
- **`j`**: select columns and summarizations to perform
    - ex: sum, count, mean, etc
- **`by`**: categorical variable to group by

Filter rows

```
df[siteid == 'Rosemount', ]
```

- goto `siteid` column == show me all rows that have Rosemount

- results:

console - for rosemount `siteid` - see precipitation, temperature…

```
> # Filter rows
> df[siteid == 'Rosemount', ]
            V1  data       date                param    siteid
     1:     40  0.00 2003-02-09        Precipitation Rosemount
     2:     86  0.00 2003-03-27        Precipitation Rosemount
     3:    132  0.00 2003-05-12        Precipitation Rosemount
     4:    178  0.84 2003-06-27        Precipitation Rosemount
     5:    224  0.00 2003-08-12        Precipitation Rosemount
    ---
14243: 655172 16.19 2015-06-25  Minimum temperature Rosemount
14244: 655218 16.45 2015-08-10  Minimum temperature Rosemount
14245: 655264 12.08 2015-09-25  Minimum temperature Rosemount
14246: 655310 -3.55 2015-11-10  Minimum temperature Rosemount
14247: 655356 -2.24 2015-12-26  Minimum temperature Rosemount
```

Select columns

```
df[, data, sideid]
```

- result:

see all of the rows , `wnr` columns data & `siteid`

```
> # Select columns
> df[, data, siteid]
          siteid  data
     1:     ACRE  0.00
     2:     ACRE  0.58
     3:     ACRE  0.00
     4:     ACRE  0.00
     5:     ACRE  0.00
    ---
655358: Walworth 12.33
655359: Walworth 18.43
655360: Walworth  5.97
655361: Walworth  5.94
655362: Walworth -8.23
```

Filter rows & select columns

```
df[siteid == 'Rosemount', data, sideid]
```

- results:

Rosemount, with just `data` & `siteid` columns

- fast

```
> # Filter rows and select columns
> df[siteid == 'Rosemount', data, siteid]
       siteid  data
    1: Rosemount  0.00
    2: Rosemount  0.00
    3: Rosemount  0.00
    4: Rosemount  0.84
    5: Rosemount  0.00
   ---
14243: Rosemount 16.19
14244: Rosemount 16.45
14245: Rosemount 12.08
14246: Rosemount -3.55
14247: Rosemount -2.24
```

Filter rows, select columns, and summarise by parameter

```
df[siteid == "Rosemount',
      mean(data, na.rm = TRUE),
      by = param]
```

- results:

```
> df[siteid == 'Rosemount',
+       mean(data, na.rm = TRUE),
+       by = param]
                   param        V1
1:         Precipitation  2.520729
2:  Maximum temperature 15.053389
3:  Minimum temperature  3.806797
```