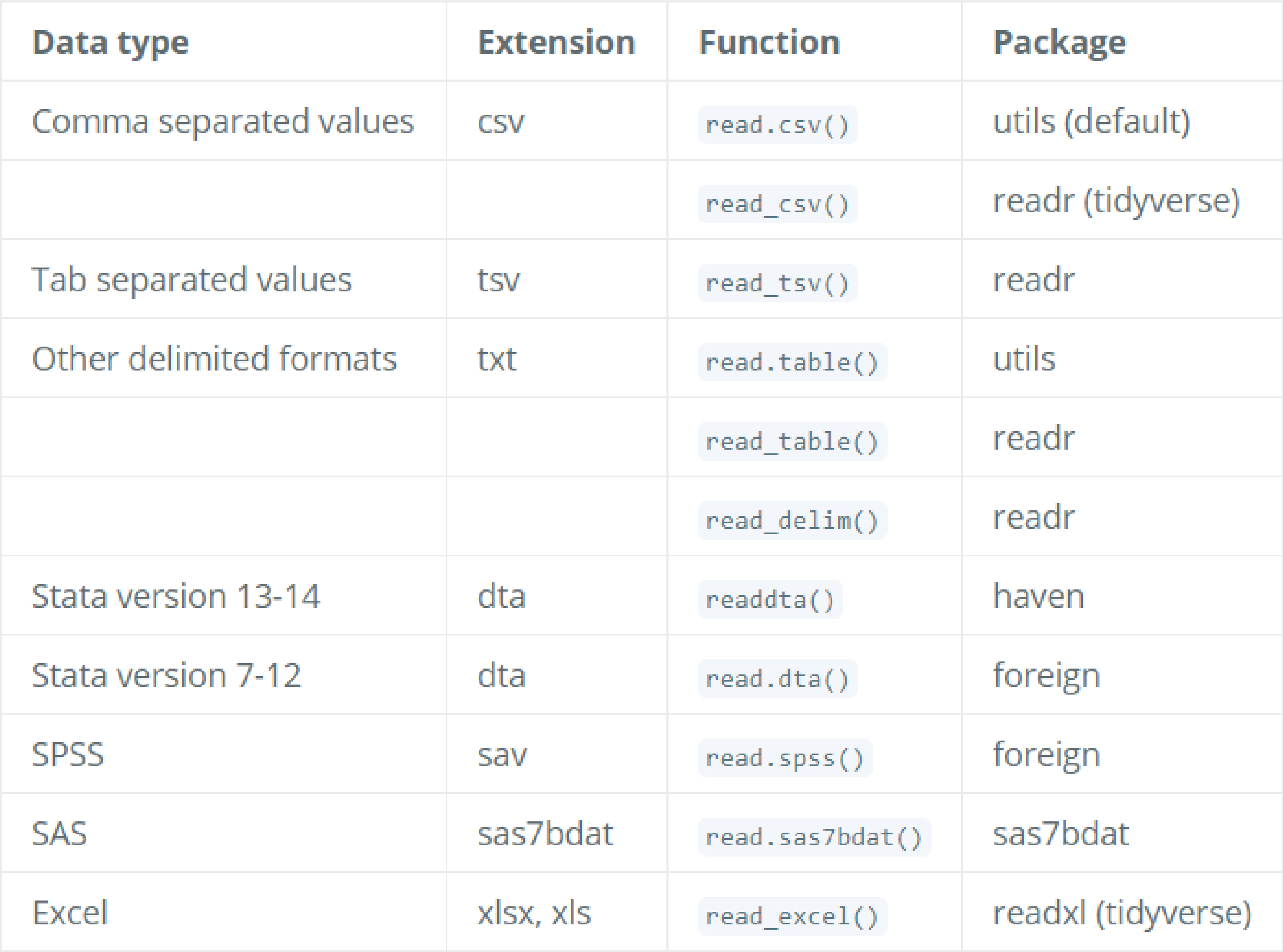
Reading & inspecting data in R

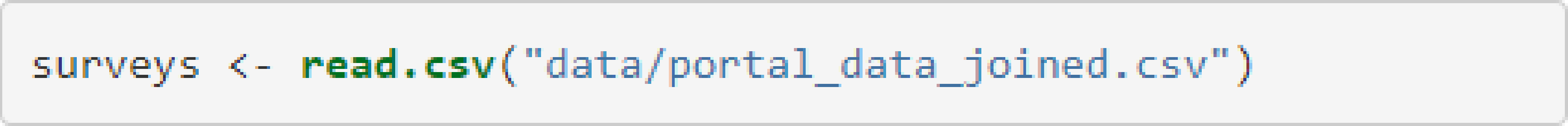
- R can read and write data from a multitude of different sources
    - Text (csv, tsv, …)
    - Stata
    - SPSS
    - SAS
    - Excel
    - …
- And using lots of different packages and functions



Data type	Extension	Function	Package
Comma separated values	csv	`read.csv()`	utils (default)
		`read_csv()`	readr (tidyverse)
Tab separated values	tsv	`read_tsv()`	readr
Other delimited formats	txt	`read.table()`	utils
		`read_table()`	readr
		`read_delim()`	readr
Stata version 13-14	dta	`readdta()`	haven
Stata version 7-12	dta	`read.dta()`	foreign
SPSS	sav	`read.spss()`	foreign
SAS	sas7bdat	`read.sas7bdat()`	sas7bdat
Excel	xlsx, xls	`read_excel()`	readxl (tidyverse)

Example

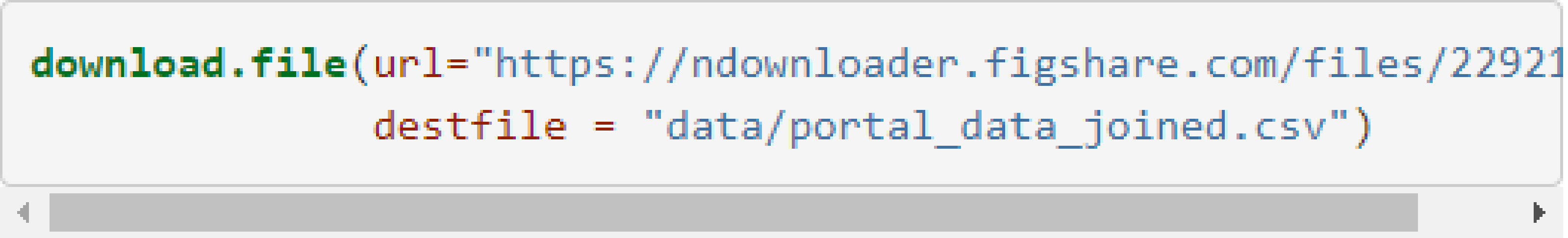
- Simple base R option



```r
surveys <- read.csv("data/portal_data_joined.csv")
```

Downloading from the web

- We can download data before reading it in to R
- A common function for this is `download.file()`
- E.g.

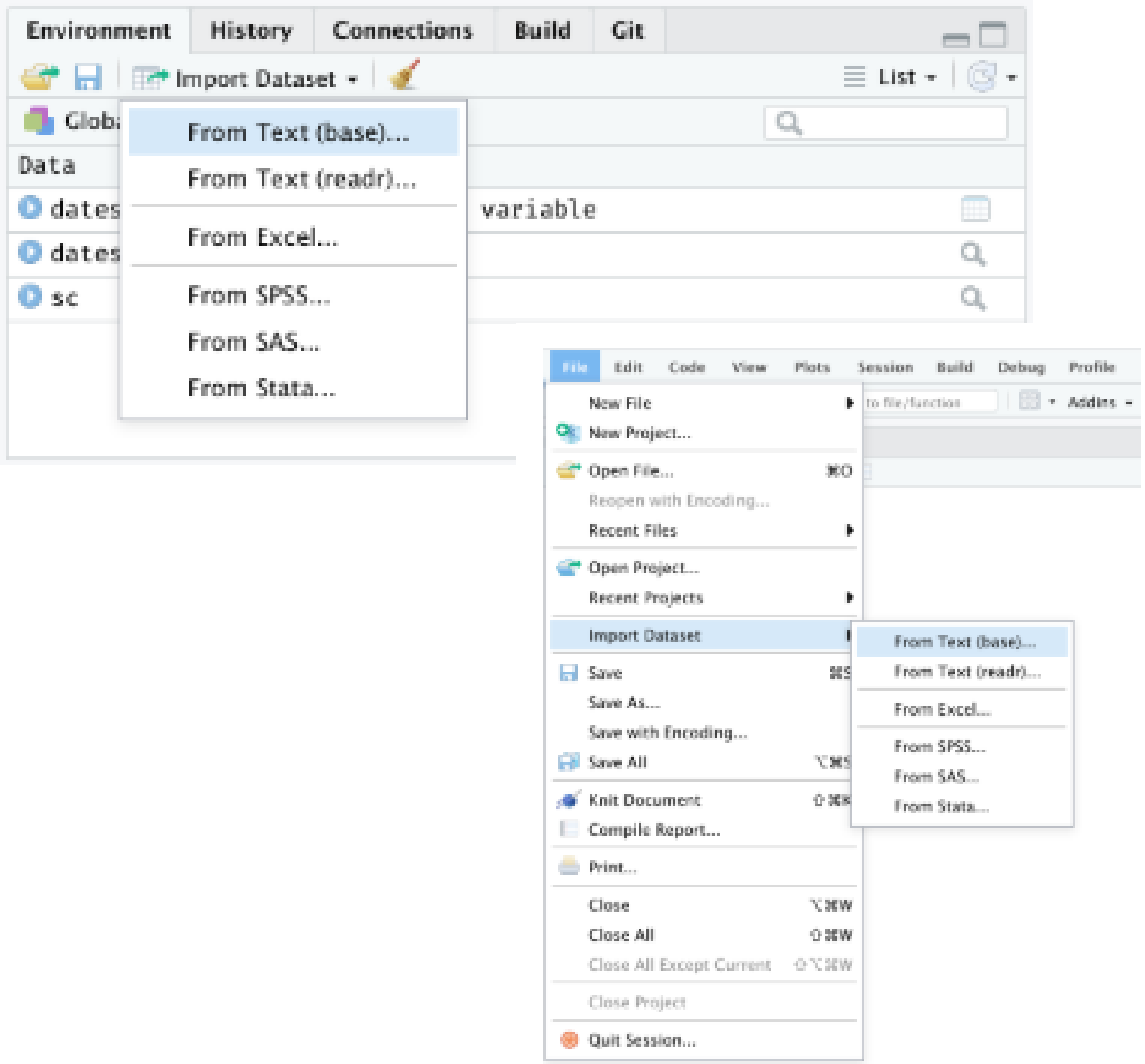


```
download.file(url="https://ndownloader.figshare.com/files/22921
              destfile = "data/portal_data_joined.csv")
```

- Can download from e.g. GitHub, figshare, googlesdrive, …

Via GUI

- 3 categories
  - Text data
  - Excel data
  - Statistical data
- Options
  - Environment pane
  - File menu



Importing using readr

- Provides support for
  - Import from the file system or a url
  - Change column data types
  - Skip or include-only columns
  - Rename the data set
  - Skip the first N rows
  - Use the header row for column names
  - Trim spaces in names
  - Change the column delimiter
  - Encoding selection
  - Select quote, escape, comment and NA identifiers

Using stringsAsFactors=FALSE

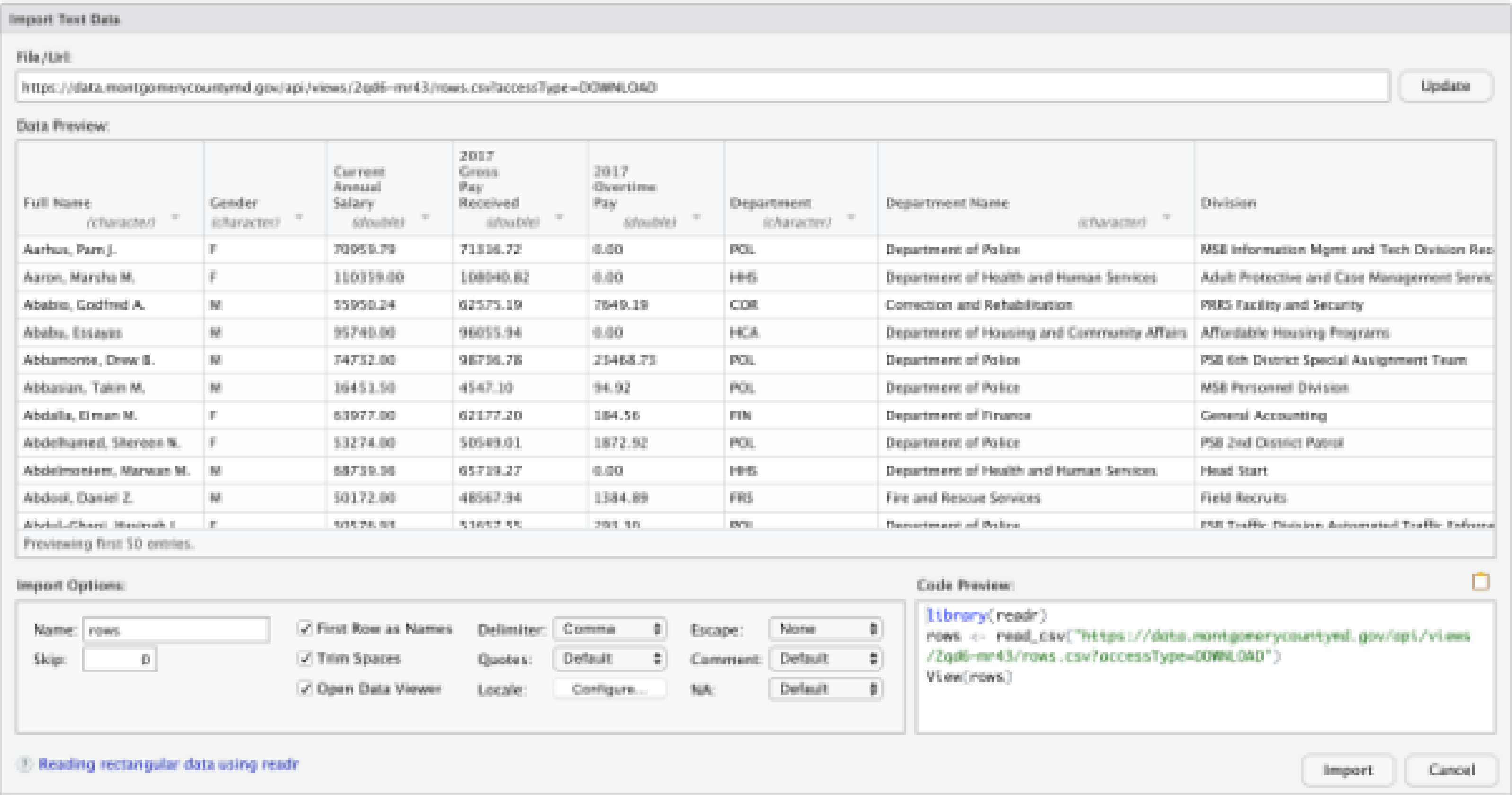
- By default, when building or importing a data frame, the columns that contain characters (i.e. text) are coerced (= converted) into factors. Depending on what you want to do with the data, you may want to keep these columns as character. To do so, read.csv() and read.table() have an argument called stringsAsFactors which can be set to FALSE.

- In most cases, it is preferable to set stringsAsFactors = FALSE when importing data and to convert as a factor only the columns that require this data type.

https://support.rstudio.com/hc/en-us/articles/218611977-Importing-Data-with-RStudio

Example

- import form data.gov (paste https://data.montgomerycountymd.gov/api/views/2qd6-mr43/rows.csv?accessType=DOWNLOAD



File/Url:

https://data.montgomerycountymd.gov/api/views/2qd6-mr43/rows.csv?accessType=DOWNLOAD

Data Preview:

Full Name (character)	Gender (character)	Current Annual Salary (double)	2017 Gross Pay Received (double)	2017 Overtime Pay (double)	Department (character)	Department Name (character)	Division
Aarhus, Pam J.	F	70959.79	71316.72	0.00	POL	Department of Police	MSB Information Mgmt and Tech Division Rec
Aaron, Marsha M.	F	110359.00	108040.82	0.00	HHS	Department of Health and Human Services	Adult Protective and Case Management Servic
Ababio, Godfred A.	M	55950.24	62575.19	7649.19	COR	Correction and Rehabilitation	PRRS Facility and Security
Ababu, Essayas	M	95740.00	96055.94	0.00	HCA	Department of Housing and Community Affairs	Affordable Housing Programs
Abbamonte, Drew B.	M	74732.00	98736.78	23468.73	POL	Department of Police	PSB 6th District Special Assignment Team
Abbasian, Takin M.	M	16451.50	4547.10	94.92	POL	Department of Police	MSB Personnel Division
Abdalla, Elman M.	F	63977.00	62177.20	184.56	FIN	Department of Finance	General Accounting
Abdelhamed, Shereen N.	F	53274.00	50549.01	1872.92	POL	Department of Police	PSB 2nd District Patrol
Abdelmoniem, Marwan M.	M	68739.36	65719.27	0.00	HHS	Department of Health and Human Services	Head Start
Abdool, Daniel Z.	M	50172.00	48567.94	1384.89	FRS	Fire and Rescue Services	Field Recruits
Abdul-Ghani, Hasinah I.	F	50576.93	51657.55	293.30	POL	Department of Police	PSB Traffic Division Automated Traffic Enforce

Previewing first 50 entries.

Import Options:

Name: rows

Skip: 0

☑ First Row as Names

☑ Trim Spaces

☑ Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

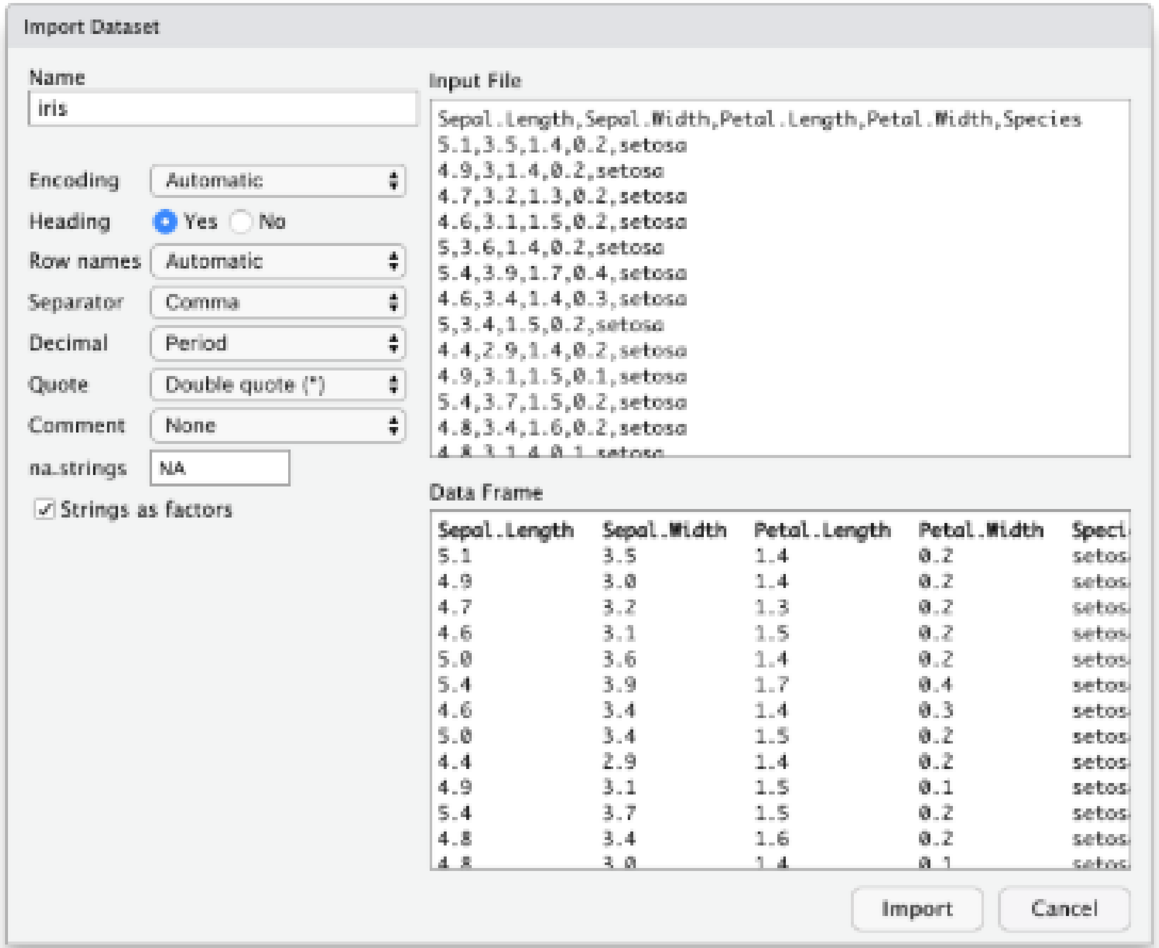
NA: Default

Code Preview:

```
library(readr)
rows <- read_csv("https://data.montgomerycountymd.gov/api/views
/2qd6-mr43/rows.csv?accessType=DOWNLOAD")
View(rows)
```

Reading rectangular data using readr

Importing data from Text files



Name: iris

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Automatic

Separator: Comma

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: NA

☑ Strings as factors

Input File

```
Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Species
5.1,3.5,1.4,0.2,setosa
4.9,3,1.4,0.2,setosa
4.7,3.2,1.3,0.2,setosa
4.6,3.1,1.5,0.2,setosa
5,3.6,1.4,0.2,setosa
5.4,3.9,1.7,0.4,setosa
4.6,3.4,1.4,0.3,setosa
5,3.4,1.5,0.2,setosa
4.4,2.9,1.4,0.2,setosa
4.9,3.1,1.5,0.1,setosa
5.4,3.7,1.5,0.2,setosa
4.8,3.4,1.6,0.2,setosa
4.8,3,1.4,0.1,setosa
```

Data Frame

```
Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Speci
5.1           3.5          1.4           0.2          setos
4.9           3.0          1.4           0.2          setos
4.7           3.2          1.3           0.2          setos
4.6           3.1          1.5           0.2          setos
5.0           3.6          1.4           0.2          setos
5.4           3.9          1.7           0.4          setos
4.6           3.4          1.4           0.3          setos
5.0           3.4          1.5           0.2          setos
4.4           2.9          1.4           0.2          setos
4.9           3.1          1.5           0.1          setos
5.4           3.7          1.5           0.2          setos
4.8           3.4          1.6           0.2          setos
4.8           3.0          1.4           0.1          setos
```

Importing data from Excel files

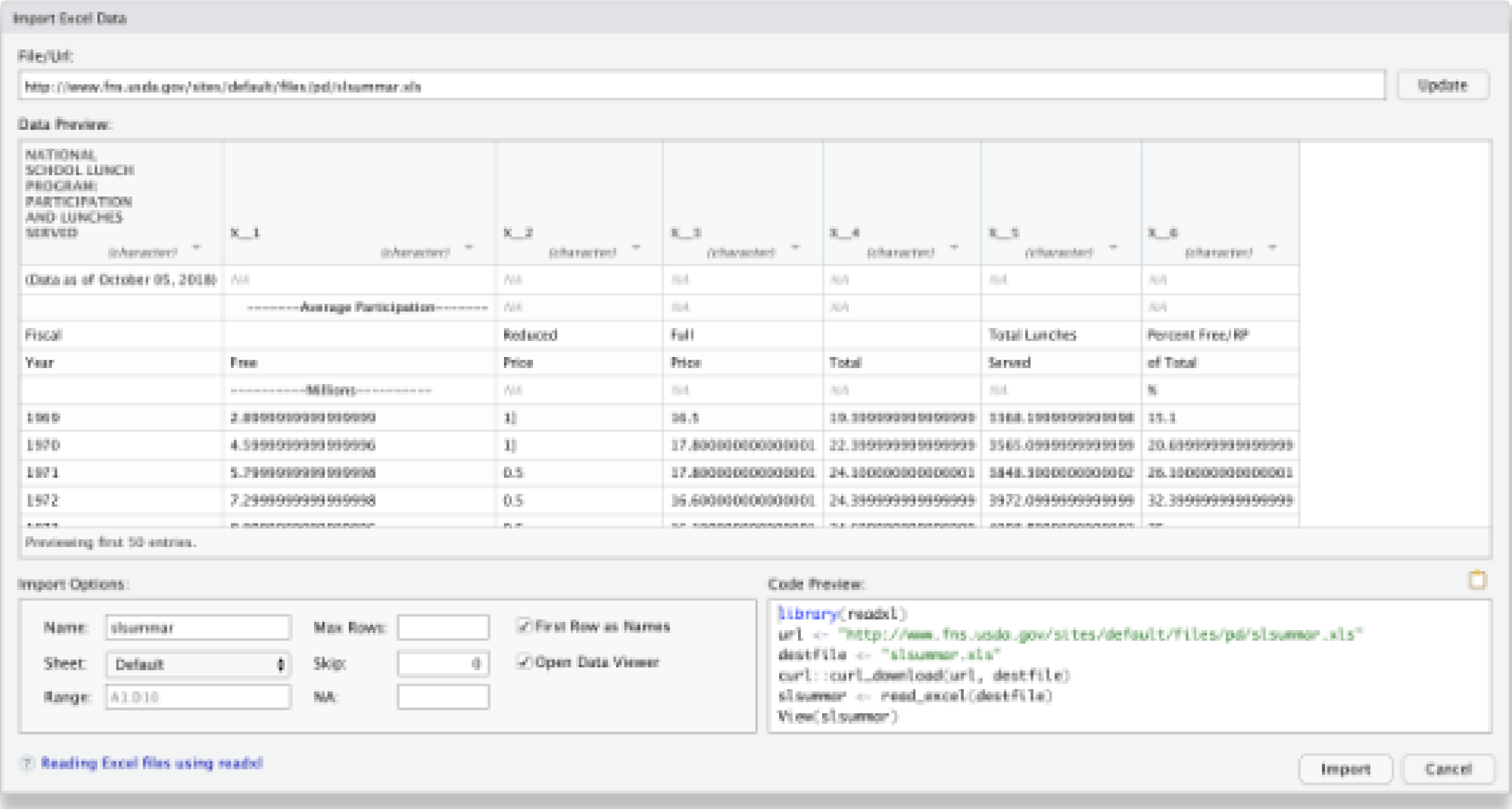
- The Excel importer provides support to:
  - Import from the file system or a url
  - Change column data types
  - Skip columns
  - Rename the data set
  - Select an specific Excel sheet
  - Skip the first N rows
  - Select NA identifiers

Example

- For example, one can import with ease an xls file from data.gov by pasting this url

http://www.fns.usda.gov/sites/default/files/pd/slsummar.xls

- selecting "Update".



Import Excel Data

File/Url:

http://www.fns.usda.gov/sites/default/files/pd/slsummar.xls

Update

Data Preview:

NATIONAL SCHOOL LUNCH PROGRAM: PARTICIPATION AND LUNCHES SERVED (character)	X__1 (character)	X__2 (character)	X__3 (character)	X__4 (character)	X__5 (character)	X__6 (character)
(Data as of October 05, 2018)	NA	NA	NA	NA	NA	NA
	--------Average Participation--------	NA	NA	NA		NA
Fiscal		Reduced	Full		Total Lunches	Percent Free/RP
Year	Free	Price	Price	Total	Served	of Total
	-----------Millions-----------	NA	NA	NA	NA	%
1969	2.8999999999999999	1/	16.5	19.399999999999999	3368.1999999999998	15.1
1970	4.5999999999999996	1/	17.800000000000001	22.399999999999999	3565.0999999999999	20.699999999999999
1971	5.7999999999999998	0.5	17.800000000000001	24.100000000000001	3848.3000000000002	26.100000000000001
1972	7.2999999999999998	0.5	16.600000000000001	24.399999999999999	3972.0999999999999	32.399999999999999

Previewing first 50 entries.

Import Options:

Name: slsummar

Max Rows:

☑ First Row as Names

Sheet: Default

Skip: 0

☑ Open Data Viewer

Range: A1:D10

NA:

Code Preview:

```
library(readxl)
url <- "http://www.fns.usda.gov/sites/default/files/pd/slsummar.xls"
destfile <- "slsummar.xls"
curl::curl_download(url, destfile)
slsummar <- read_excel(destfile)
View(slsummar)
```

- We can clean this up by
  - skipping 6 rows from this file
  - unchecking the "First Row as Names" checkbox.
- The file is looking better but some columns are being displayed as strings when they are clearly numerical data.
- We can fix this by selecting "numeric" from the column dropdown.
- The final step is to click "Import" to run the code under "Code Preview" and import the data into RStudio, the final result should look as follows:

Inspecting data structure

- Once data is read-in there are several way to view and interrogate them

- The most simple is to type the name of the variable
- Can you see any problems with this?

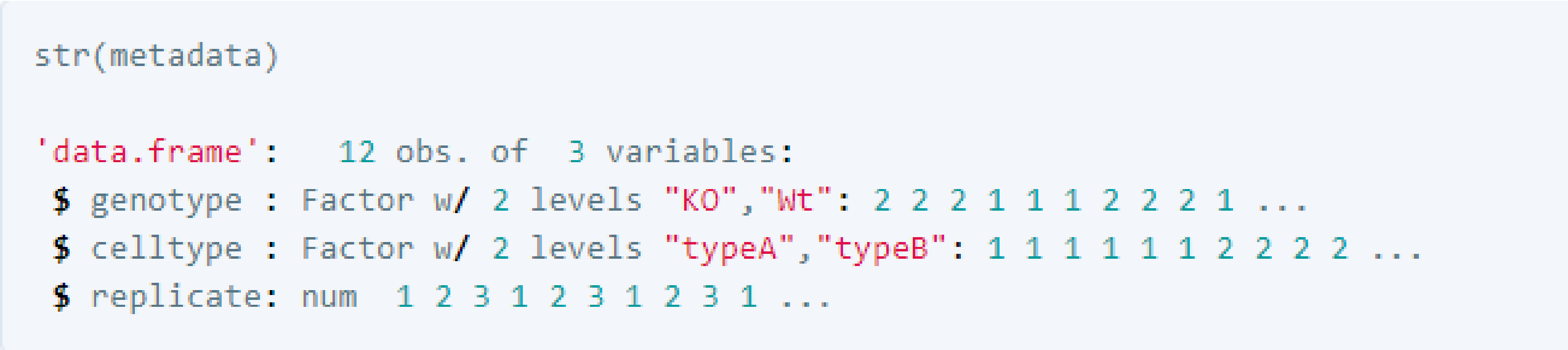
Other options

- View the top of the data
- head(<name>)

- Equivalently we can use
- tail(<name>)

Compact display of contents

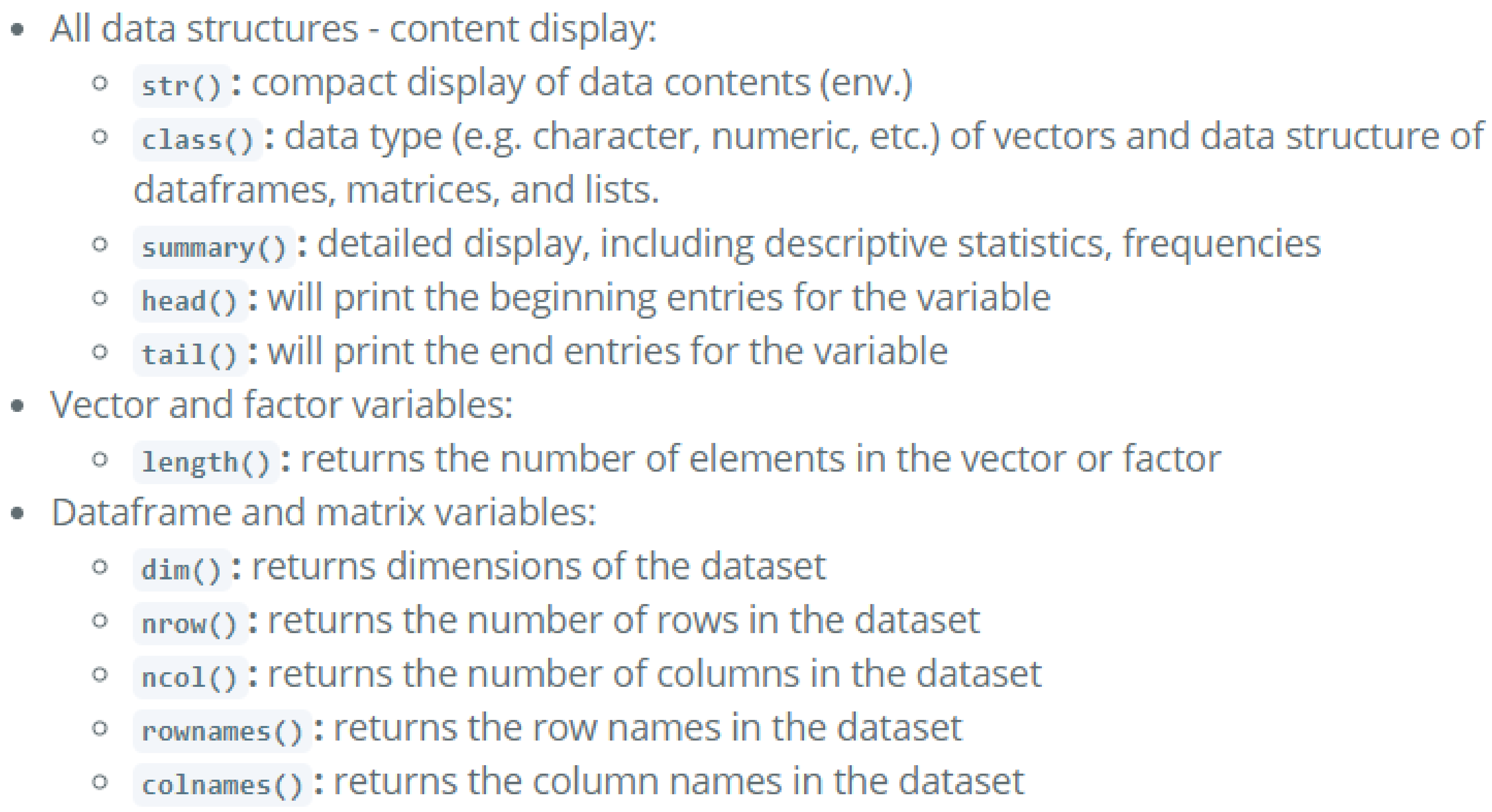
- The head and tail are limited in terms of information
- A good overall summary in base R is
- str(<name>)



```
str(metadata)

'data.frame':   12 obs. of  3 variables:
 $ genotype : Factor w/ 2 levels "KO","Wt": 2 2 2 1 1 1 2 2 2 1 ...
 $ celltype : Factor w/ 2 levels "typeA","typeB": 1 1 1 1 1 1 2 2 2 2 ...
 $ replicate: num  1 2 3 1 2 3 1 2 3 1 ...
```

List of some data inspection functions



- All data structures - content display:
  - `str()`: compact display of data contents (env.)
  - `class()`: data type (e.g. character, numeric, etc.) of vectors and data structure of dataframes, matrices, and lists.
  - `summary()`: detailed display, including descriptive statistics, frequencies
  - `head()`: will print the beginning entries for the variable
  - `tail()`: will print the end entries for the variable
- Vector and factor variables:
  - `length()`: returns the number of elements in the vector or factor
- Dataframe and matrix variables:
  - `dim()`: returns dimensions of the dataset
  - `nrow()`: returns the number of rows in the dataset
  - `ncol()`: returns the number of columns in the dataset
  - `rownames()`: returns the row names in the dataset
  - `colnames()`: returns the column names in the dataset

Viewing data

- You can view the data (when data frame, matrix etc) in the data panel in several ways:

```
View(<name>)
Edit(<name>)
```

- Left click with the mouse on the name of the variable in the Environment tab
- Hover over the variable name in the console or scripts and press F2