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)

raw -> technically correct data

- `read.table` and its cousins (`utils` package)
  - `read.csv:` *comma* separated values with *period* as decimal separator.
  - `read.csv2:` *semicolon* separated values with *comma* as decimal separator.
  - `read.delim:` *tab-delimited* files with *period* as decimal separator.
  - `read.delim2:` *tab-delimited* files with *comma* as decimal separator.
  - `read.fwf:` data with a predetermined number of bytes per column.

Example

- Simple base R option

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

Downloading from the web

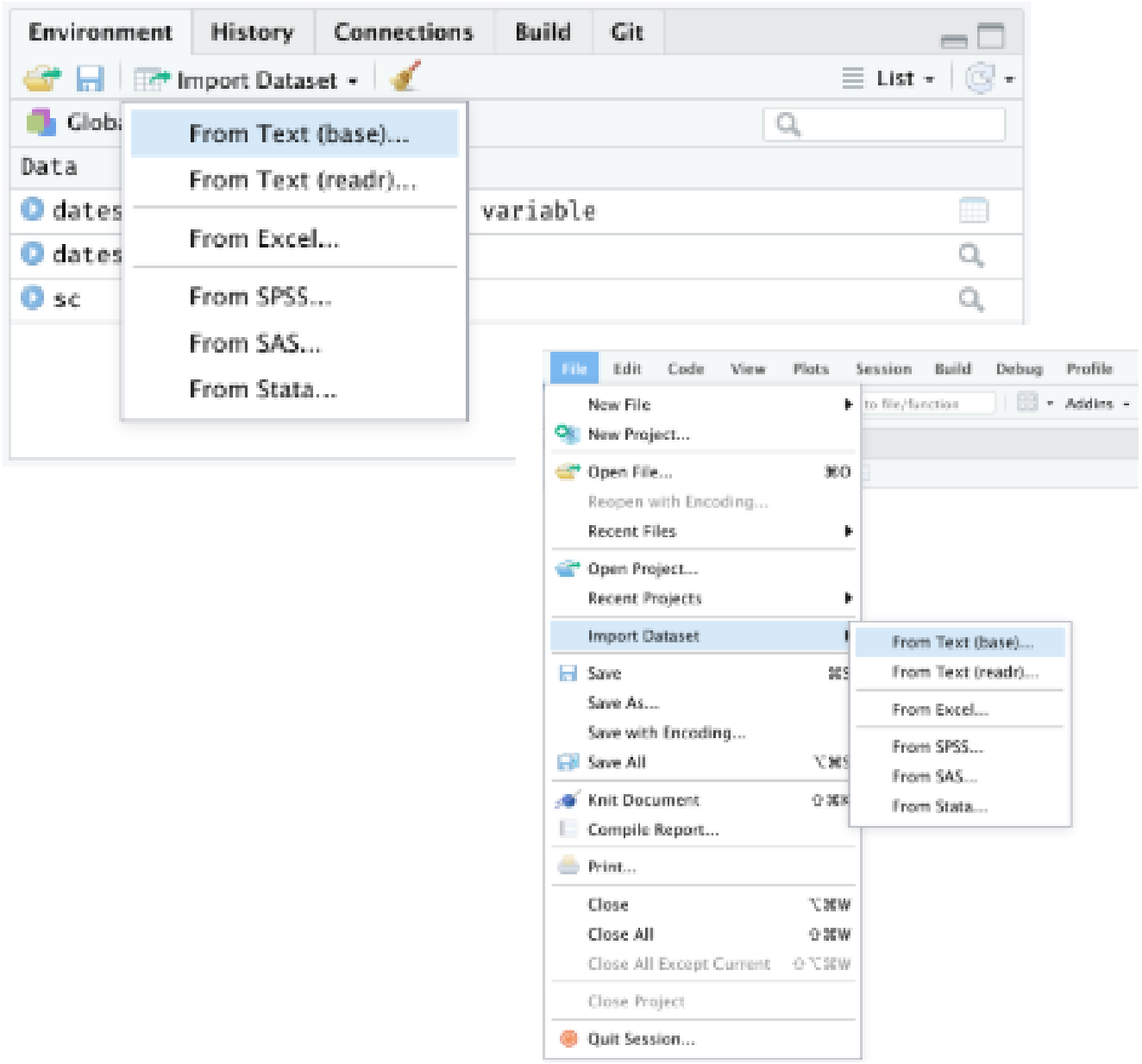
- We can download data before reading it in to R
- A common function for this is `download.file()`
- E.g.

```
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

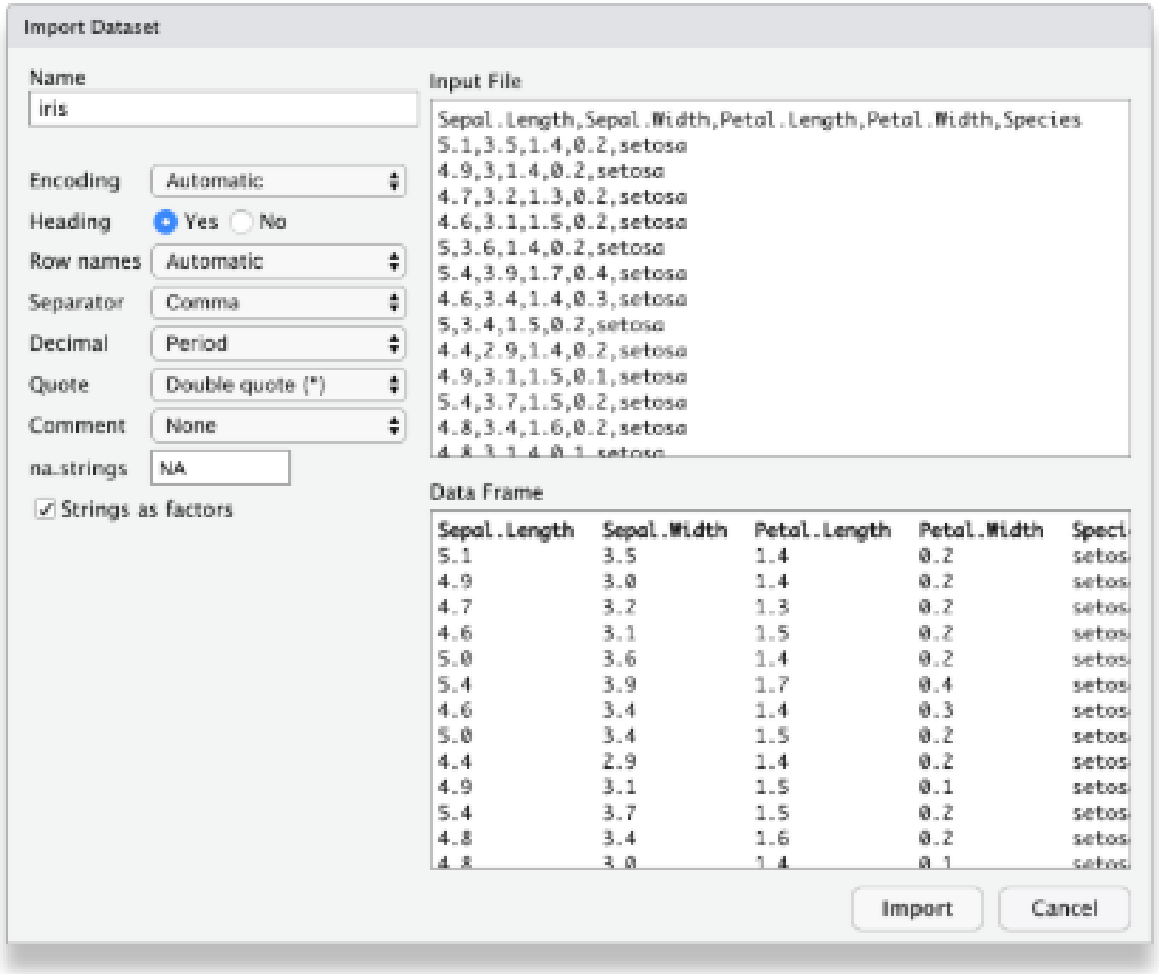
- Provides support for
  - Import from the file system or a url
  - Change column data types
  - Skip or include-only columns
  - Rename the data set
  - Skip the first N rows
  - Use the header row for column names
  - Trim spaces in names
  - Change the column delimiter
  - Encoding selection
  - Select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

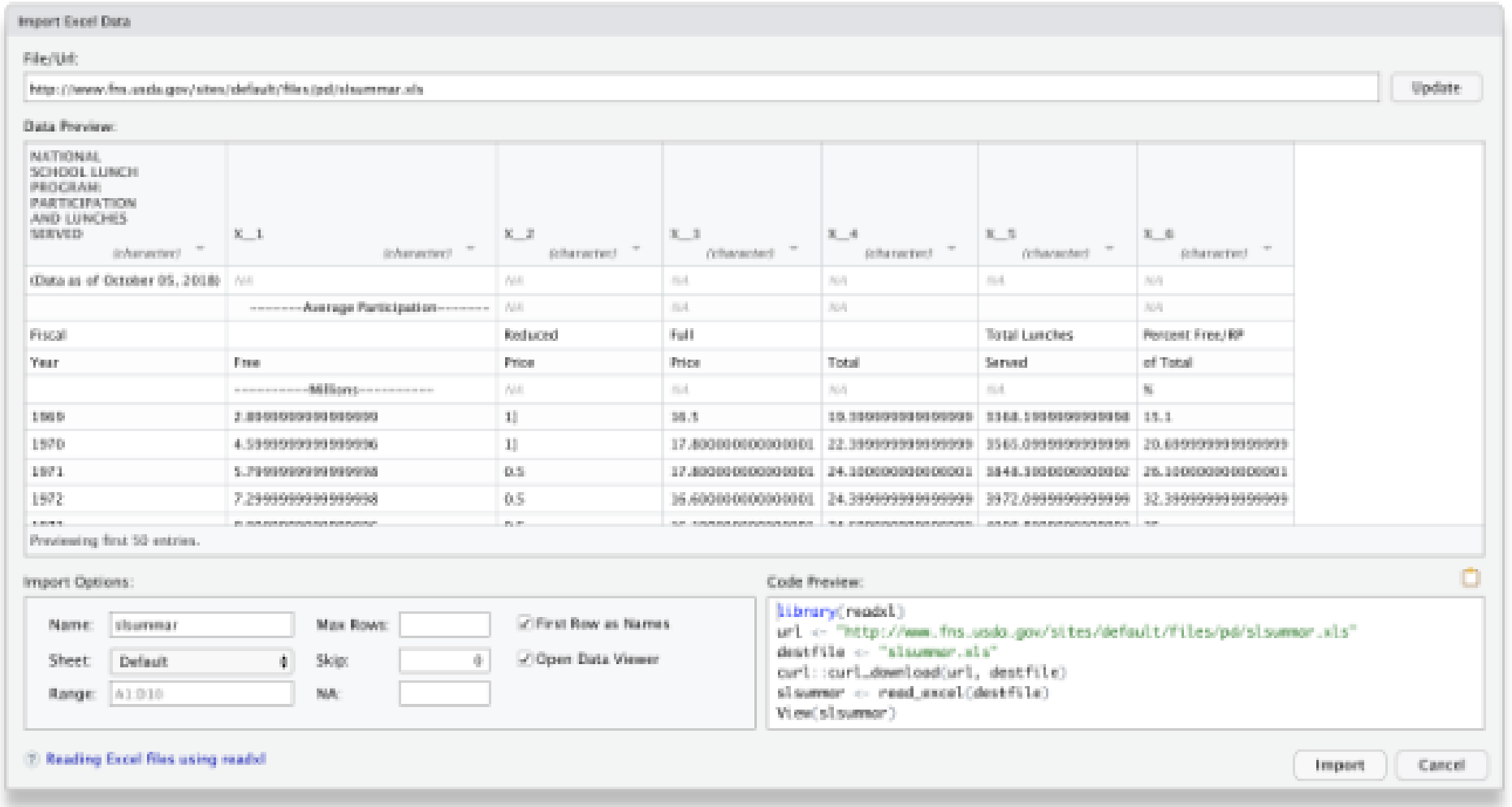
- In most cases, it is preferable to set stringsAsFactors = FALSE when importing data and to convert as a factor only the columns that require this data type.

Importing data from Text files



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


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)
--------Average Participation--------
Fiscal
Year
Free
Reduced Price
Full Price
Total
Total Lunches Served
Percent Free/RP of Total
----------Millions----------
%
1969 2.8999999999999999 1) 18.5 19.399999999999999 3168.1999999999998 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
Sheet: Default
Range: A1:D10
Max Rows:
Skip: 0
NA:
First Row as Names
Open Data Viewer
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)
Reading Excel files using readxl
Import
Cancel

- 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:

.RData, .Rda, .Rds

- Store R objects with properties
- .Rda is just a short name for .Rdata
  - `save(), load(), attach()`,…
- .Rds use to restore R object with different name
  - `readRDS(), saveRDS()`,…

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

Writing to Excel

- Format, styling and editing inside of R
- `library(openxlsx)`
  - `l <- list(iris = iris, mtcars = mtcars, chickwts = chickwts, quakes = quakes)`
  - `openxlsx::write.xlsx(l, file="datasets.xlsx")`
- Much more fancy features
- E.g.
- https://cran.r-project.org/web/packages/openxlsx/vignettes/formatting.pdf

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Tell me what you

Cut | Copy | Format Painter | Paste — Clipboard

Calibri 11 — Font

Wrap Text | Merge & Center — Alignment

A	Date	Logical	Curren	Accoun	hLink	Percen	TinyNu
	#########	TRUE	-$2.00	-£ 2.00	https://CF	-100.00%	5.66E-10
	#########	FALSE	-$1.00	-£ 1.00	https://CF	-50.00%	8.55E-10
	#########	TRUE	$0.00	£ -	https://CF	0.00%	5.80E-10
	#########	TRUE	$1.00	£ 1.00	https://CF	50.00%	1.52E-10
	#########	FALSE	$2.00	£ 2.00	https://CF	100.00%	3.75E-10

Date	Logical	Curren	Accoun	hLink	Percen	TinyNu
#########	TRUE	-$2.00	-£ 2.00	https://CF	-100.00%	5.66E-10
#########	FALSE	-$1.00	-£ 1.00	https://CF	-50.00%	8.55E-10
#########	TRUE	$0.00	£ -	https://CF	0.00%	5.80E-10
#########	TRUE	$1.00	£ 1.00	https://CF	50.00%	1.52E-10
#########	FALSE	$2.00	£ 2.00	https://CF	100.00%	3.75E-10

Date	Logical	Curren	Accoun	hLink	Percen	TinyNu…ber
#########	TRUE	-$2.00	-£ 2.00	https://CF	-100.00%	5.66E-10
#########	FALSE	-$1.00	-£ 1.00	https://CF	-50.00%	8.55E-10
#########	TRUE	$0.00	£ -	https://CF	0.00%	5.80E-10
#########	TRUE	$1.00	£ 1.00	https://CF	50.00%	1.52E-10
#########	FALSE	$2.00	£ 2.00	https://CF	100.00%	3.75E-10

Date	Logical	Curren	Accoun	hLink	Percen	TinyNu
#########	TRUE	-$2.00	-£ 2.00	https://CF	-100.00%	5.66E-10
#########	FALSE	-$1.00	-£ 1.00	https://CF	-50.00%	8.55E-10
#########	TRUE	$0.00	£ -	https://CF	0.00%	5.80E-10
#########	TRUE	$1.00	£ 1.00	https://CF	50.00%	1.52E-10
#########	FALSE	$2.00	£ 2.00	https://CF	100.00%	3.75E-10

writeData auto-formatting | writeDataTable