Intro to R

Data Input/Output

Outline

- Part 1: reading CSV file, common new user mistakes in data reading, checking for problems in the read data
- Part 2: data input overview, working directories, relative vs. absolute paths, reading XLSX file (Excel file), other data inputs
- Part 3: writing CSV file
- Part 4: reading and saving R objects

Data We Use

- Everything we do in class will be using real publicly available data - there are few ‘toy’ example datasets and ‘simulated’ data
- Baltimore Open Data and Data.gov will be sources for the first few days
- We have also added functionality to load these datasets directly in the `jhur` package

Data Input

- 'Reading in' data is the first step of any real project/analysis
- R can read almost any file format, especially via add-on packages
- We are going to focus on simple delimited files first
    - comma separated (e.g. '.csv')
    - tab delimited (e.g. '.txt')
    - Microsoft Excel (e.g. '.xlsx')

Data Input

Youth Tobacco Survey (YTS) dataset:

“The YTS was developed to provide states with comprehensive data on both middle school and high school students regarding tobacco use, exposure to environmental tobacco smoke, smoking cessation, school curriculum, minors’ ability to purchase or otherwise obtain tobacco products, knowledge and attitudes about tobacco, and familiarity with pro-tobacco and anti-tobacco media messages.”

- Check out the data at: https://catalog.data.gov/dataset/youth-tobacco-survey-yts-data

Data Input: Dataset Location

Dataset is located at
http://jhudatascience.org/intro_to_r/data/Youth_Tobacco_Survey_YTS_Data.csv

- Download data by clicking the above link
    - Safari - if a file loads in your browser, choose File –> Save As, select, Format “Page Source” and save

Data Input: Read in Directly

```
# load library `readr` that contains function `read_csv`
library(readr)
dat = read_csv("http://jhudatascience.org/intro_to_r/data/Youth_Tobacco_Survey

# `head` displays first few rows of a data frame
head(dat, 5)
```

```
# A tibble: 5 x 31
   YEAR LocationAbbr LocationDesc TopicType  TopicDesc  MeasureDesc   DataSour
  <dbl> <chr>        <chr>        <chr>      <chr>      <chr>         <chr>
1  2015 AZ           Arizona      Tobacco U… Cessation… Percent of C… YTS
2  2015 AZ           Arizona      Tobacco U… Cessation… Percent of C… YTS
3  2015 AZ           Arizona      Tobacco U… Cessation… Percent of C… YTS
4  2015 AZ           Arizona      Tobacco U… Cessation… Quit Attempt… YTS
5  2015 AZ           Arizona      Tobacco U… Cessation… Quit Attempt… YTS
# … with 24 more variables: Response <chr>, Data_Value_Unit <chr>,
#   Data_Value_Type <chr>, Data_Value <dbl>, Data_Value_Footnote_Symbol <chr>,
#   Data_Value_Footnote <chr>, Data_Value_Std_Err <dbl>,
#   Low_Confidence_Limit <dbl>, High_Confidence_Limit <dbl>, Sample_Size <dbl>
#   Gender <chr>, Race <chr>, Age <chr>, Education <chr>, GeoLocation <chr>,
#   TopicTypeId <chr>, TopicId <chr>, MeasureId <chr>, StratificationID1 <chr>
#   StratificationID2 <chr>, StratificationID3 <chr>, StratificationID4 <chr>,
#   SubMeasureID <chr>, DisplayOrder <dbl>
```

Data Input: Read in Directly

So what is going on “behind the scenes”?

`read_csv()` parses a “flat” text file (.csv) and turns it into a **tibble** – a rectangular data frame, where data are split into rows and columns

- First, a flat file is parsed into a rectangular matrix of strings
- Second, the type of each column is determined (heuristic-based guess)

Data Input: Read in Directly

`read_csv()` needs path to your file, will return a tibble

```
read_csv(file, col_names = TRUE, col_types = NULL,
  locale = default_locale(), na = c("", "NA"),
  quoted_na = TRUE, quote = "\"", comment = "", trim_ws = TRUE,
  skip = 0, n_max = Inf, guess_max = min(1000, n_max),
  progress = show_progress(), skip_empty_rows = TRUE
)
```

- `file` is the path to your file, in quotes
- can be path in your local computer – absolute file path or relative file path
- can be path to a file on a website

```
## Examples

dat = read_csv("/Users/martakaras/Downloads/Youth_Tobacco_Survey_YTS_Data.csv"

dat = read_csv("Youth_Tobacco_Survey_YTS_Data.csv")

dat = read_csv("www.someurl.com/table1.csv")
```

Data Input: Read in Directly

`read_csv()` is a special case of `read_delim()` – a general function to read a delimited file into a data frame

`read_delim()` needs path to your file and fileds delimiter, will return a tibble

```
read_delim(file, delim, quote = "\"", escape_backslash = FALSE,
  escape_double = TRUE, col_names = TRUE,  col_types = NULL,
  locale = default_locale(),na = c("", "NA"),  quoted_na = TRUE,
  comment = "", trim_ws = FALSE, skip = 0,
  n_max = Inf,  guess_max = min(1000, n_max),
  progress = show_progress(), skip_empty_rows = TRUE
)
```

- `file` is the path to your file, in quotes
- `delim` is what separates the fields within a record

```
## Examples
dat = read_delim("Youth_Tobacco_Survey_YTS_Data.csv", delim = ",")

dat = read_delim("www.someurl.com/table1.txt", delim = "\t")
```

Data Input: Read in Directly From File Path

```
dat = read_csv("../data/Youth_Tobacco_Survey_YTS_Data.csv")
```

```
── Column specification ────────────────────────────────────────────────────────
cols(
  .default = col_character(),
  YEAR = col_double(),
  Data_Value = col_double(),
  Data_Value_Std_Err = col_double(),
  Low_Confidence_Limit = col_double(),
  High_Confidence_Limit = col_double(),
  Sample_Size = col_double(),
  DisplayOrder = col_double()
)
ℹ Use `spec()` for the full column specifications.
```

The data is now successfully read into your R workspace. Colum specification of first few columns is printed to the console.

Common new user mistakes we have seen

1. **Working directory problems: trying to read files that R "can't find"**
   - Path misspecification
2. Typos (R is **case sensitive**, `x` and `X` are different)
   - RStudio helps with "tab completion"
3. Data type problems (is that a string or a number?)
4. Open ended quotes, parentheses, and brackets
5. Different versions of software

Data Input: Checking for problems

- The `spec()` and `problems()` functions show you the specification of how the data was read in.

```
problems(dat)
```

```
[1] row      col      expected actual
<0 rows> (or 0-length row.names)
```

```
spec(dat)
```

```
cols(
  YEAR = col_double(),
  LocationAbbr = col_character(),
  LocationDesc = col_character(),
  TopicType = col_character(),
  TopicDesc = col_character(),
  MeasureDesc = col_character(),
  DataSource = col_character(),
  Response = col_character(),
  Data_Value_Unit = col_character(),
  Data_Value_Type = col_character(),
  Data_Value = col_double(),
  Data_Value_Footnote_Symbol = col_character(),
  Data_Value_Footnote = col_character(),
  Data_Value_Std_Err = col_double(),
  Low_Confidence_Limit = col_double(),
  High_Confidence_Limit = col_double(),
  Sample_Size = col_double(),
  Gender = col_character(),
```

Data Input: Checking for problems

The `stop_for_problems()` function will stop if your data had any problem when reading in (even if that problem did not cause the data reading to fail).

- Particularly useful to put after the data reading code e.g. in some automated R script that should not proceed in case some data "weirdness" occurred.

```
stop_for_problems(dat)
```

Help

For any function, you can write `?FUNCTION_NAME`, or `help("FUNCTION_NAME")` to look at the help file:

```
?read_delim
help("read_delim")
```

Data Input: Read in From RStudio Toolbar

R Studio features some nice “drop-down” support, where you can run some tasks by selecting them from the toolbar.

For example, you can easily import text datasets using the `File --> Import Dataset --> From Text (readr)` command. Selecting this will bring up a new screen that lets you specify the formatting of your text file.

After importing a datatset, you get (printed in the R console) the corresponding R command that you can enter in the console if you want to re-import data.

Data Input: base R

There are also data importing functions provided in base R (rather than the `readr` package), like `read.delim()` and `read.csv()`.

These functions have slightly different syntax for reading in data (e.g. `header` argument).

However, while many online resources use the base R tools, the latest version of RStudio switched to use these new `readr` data import tools, so we will use them in the class for slides. They are also up to two times faster for reading in large datasets, and have a progress bar which is nice.

But you can use whatever function you feel more comfortable with.

Revision

- Data importing functions provided in base R: `read.delim()`, `read.csv()`
- Modern, improved tools from `readr` R package: `read_delim()`, `read_csv()`
  - needs a file path to be provided
  - parses the file into rows/columns, determines column type
  - returns a data frame
- Some functions to look at a data frame:
  - `head()` shows first few rows
  - `spec()` gives specification of column types

Data input: other file types

- From `readr` package:
  - `read_delim()`: general delimited files
  - `read_csv()`: comma separated (CSV) files
  - `read_tsv()`: tab separated files
  - others
- For reading Excel files, you can do one of:
  - open in Excel, “Save as” a sheet as a .csv file, and open using `read_csv()`
  - use `read_excel()` function from `readxl` package
  - use other packages: `xlsx`, `openxlsx`
- `haven` package has functions to read SAS, SPSS, Stata formats
- `sas7bdat` has functions to read SAS formats

Lab Part 1

Lab file: http://jhudatascience.org//intro_to_r/Data_IO/lab/Data_IO_Lab.Rmd

Website

Working Directories

Working directory is a directory that R assumes “you are working in”.

“Setting working directory” means specifying the path to the directory.

```
# get the working directory
getwd()

# set the working directory
setwd("/Users/martakaras/Desktop")
```

R uses working directory as a starting place when searching for files.

Working Directories

R uses working directory as a starting place when searching for files:

- if you use `read_csv("Bike_Lanes_Long.csv")`, R assumes that the file is **in** the working directory
- if you use `read_csv("data/Bike_Lanes_Long.csv")`, R assumes that `data` directory is **in** the working directory
- if you use an absolute path,
  e.g. `read_csv("/Users/martakaras/data/Bike_Lanes_Long.csv")`, the working directory information is not used

Data Output

While its nice to be able to read in a variety of data formats, it's equally important to be able to output data somewhere.

The `readr` package provides data exporting functions which have the pattern `write_*`:

- `write_csv()`,
- `write_delim()`, others.

From `write_csv()` documentation:

```
write_csv(x, file,
  na = "NA", append = FALSE,
  col_names = !append, quote_escape = "double",
  eol = "\n", path = deprecated()
)
```

Data Output

`x`: data frame you want to write

`file`: file path where you want to R object written; it can be:

- an absolute path,
- a relative path (relative to your working directory),
- a file name only (which writes the file to your working directory)

```
# Examples

write_csv(dat, file = "YouthTobacco_newNames.csv")

write_delim(dat, file = "YouthTobacco_newNames.csv", delim = ",")
```

R binary file

`.rds` is an extension for R native file format

`write_rds()` and `read_rds()` from `readr` package can be used to write/read a single R variable to/from file

```
# write a variable: a data frame "dat"
write_rds(dat, file = "yts_dataset.rds")

# write a variable: vector "x"
x <- c(1,3,3)
write_rds(x, file = "my_vector.rds")

# read a variable from file and assign to a new variable named "y"
x2 <- read_rds("my_vector.rds")
x2
```

```
[1] 1 3 3
```

Lab Part 2

Lab file: http://jhudatascience.org//intro_to_r/Data_IO/lab/Data_IO_Lab.Rmd

Website