Spreadsheets

Jim Harner

6/7/2019

The `readxl` package is part of the `tidyverse`. It is used for importing tabular Excel files (`.xls` and `.xlsx`) into R as tibbles.

```
library("readxl", warn.conflicts = F)
```

2.3 Spreadsheets

Spreadsheets are widely used as a storage option. Microsoft Excel is commonly used spreadsheet software. The corresponding spreadsheet format is Microsoft Excel workbook.

Spreadsheet formats

Previously Excel workbooks used a binary format, XLS, which was difficult to decode. Now Excel workbooks are stored in an XML-based format called Open Office XML (OOXML).

The main difference between these two file extensions is that the XLS is created on the version of Excel prior to 2007 while XLSX is created on the version of Excel 2007 and onward. XLS is a binary format while that XLSX is Open XML format.

Open Office Calc uses an XML-based standard format called Open Document Format (ODF).

Neither OOXML or ODF are good for storing data since a lot of data is stored in the spreadsheet about how to display the data, how to calculate cells, etc. This information is not relevant to the actual data. Also, the number of rows and columns is limited, so Excel is not useful for "big data."

Spreadsheet software

Spreadsheets displays a data set in a rectangular grid of cells. It has the benefits of fixed-width format text files. Spreadsheets have tools to manipulate data, but they are limited. Formulas can also be used.

The point-and-click interface of spreadsheets does not allow steps to be recorded although macros are available.

The function `read_excel` in the R package `readxl` can read both Excel 97–2007 files and Excel 2007+ files with the `read_xls` and `read_xlsx` functions, respectively. The `read_excel` auto detect the format from the file extension.

```
nemo_xls <- read_excel("pointnemotemp.xlsx", col_names = c("date", "temp"),
                       col_types = c("date", "numeric"))
nemo_xls
```

```
## # A tibble: 93 x 2
##    date                 temp
##    <dttm>              <dbl>
##  1 1994-01-16 00:00:00  279.
##  2 1994-02-16 00:00:00  280
##  3 1994-03-16 00:00:00  279.
##  4 1994-04-16 00:00:00  279.
##  5 1994-05-16 00:00:00  278.
##  6 1994-06-16 00:00:00  276.
##  7 1994-07-16 00:00:00  276.
##  8 1994-08-16 00:00:00  276.
##  9 1994-09-16 00:00:00  276.
## 10 1994-10-16 00:00:00  277.
## # ... with 83 more rows
```

Specific Excel sheets can be read with the `sheet` argument, either by sheet name or by sheet position. Cell ranges can be read with the `range` argument. See the documentation for other options.

Multiple files cannot be easily handled. Metadata and file names differentiate them and thus programming is difficult. This can be handled, but relational databases are more efficient.