DataSci Cleaning Data Lecture Notes – Reading Local Flat Files

Flat Files

1. Txt files
2. CSV Files
3. Tab delimitated files

Read with

1. Read.table()
   1. Most robust, so most used
   2. Requires more parameters than other read commands
   3. Reads into RAM, so large files can be problematic
   4. Parameters: *file, header, sep, row.names, nrows*
   5. Other Parameters:
      1. Quote: tells R if there are quoted values, quote=”” means no quotes
         1. VERY useful for messy data
      2. Na.strings – sets the character that represents a missing value
      3. Nrows – number of rows to read of the file, default is all
      4. Skip – number of lines to skip before starting to read
2. Read.csv(), read.csv2()
3. Read.xlsx
   1. Excel is the most popular format for data
   2. Has its own package library(xlsx)
   3. Parameters = *sheetIndex, header*
   4. Parameters = *colIndex, rowIndex* allows you to subset the .xlsx file as you read it into R
4. Write.xlsx
   1. Writes the findings to Excel
5. Write.xlsx2 does the same, but is faster
   1. Can have issues when subsets are read in to R
6. Other packages of use for reading in Excel
   1. XLConnect has more options