Data Cleaning and Data Loading into R

Bruce D. Marron
Project No. 2018NWDS006

August 18, 2018

Data Clean I

Output of datasets from Excel often contain spurious quotes, misplaced commas, non-ASCII characters, $ characters, and % characters. The dataset needs to be cleaned to fit the format of a simple comma-separated file. Simple comma separated files are kept as proprietary datasets. Note that there are Linux/Unix vs. Microsoft issues with regards to end-of-line formats as well as with character set encodings in text files:

- Linux/Unix uses (ASCII octal 012; LF) while Microsoft uses (ASCII octal 012 + octal 015; CRLF)
- Linux/Unix uses UTF-8 ASCII; Microsoft uses WINDOWS-1252 (a non-ISO extended-ASCII)
- It is possible to export UTF-8 encoded text file from Excel

See this article https://donatstudios.com/CSV-An-Encoding-Nightmare.

Set the ASCII character set and the end-of-line format

```
file ROIs_FOR_BRUCE_07_28_2018.csv
dos2unix -idu ROIs_FOR_BRUCE_07_28_2018.csv

iconv -f WINDOWS-1252 -t UTF-8 ROIs_FOR_BRUCE_07_28_2018.csv -o ROI.csv \
&& dos2unix ROI.csv

file ROI.csv
unix2dos -idu ROI.csv
```

Set field separators correctly and character clean a .csv file

The code below ensures that a copy of the original file (.csv) is kept and new, character cleaned (CC) version is created (CC_.csv)

1. Use 'awk' to
   - remove commas inside quotes (in titles or in numbers) then remove the quotes
   - remove $ and % characters
2. Use 'tr' to
   - remove all non-printable ASCII characters (garbage characters== !(octal 11-15 || 40-176)).
   - NB. 'tr' uses backslash to denote an octal number.

```
awk -F'"' -v OFS='' '{ for (i=2; i<=NF; i+=2) gsub(",", "", $i) }1' ROI.csv \
> CC_ROI.csv

awk 'BEGIN{FS=","} ; {gsub(/\$|%/,"",$0)}1' CC_ROI.csv > tmp.csv \
&& mv tmp.csv CC_ROI.csv
```

```
tr -cd '\11-\15\40-\176' < CC_ROI.csv > temp.csv \
&& mv temp.csv CC_ROI.csv
```

Load data into R

Assume data have been put in a file, "data".

```
data_path = file.path(getwd(), "data")
datafile = file.path(data_path, "2018RG005_data.csv")
data <- readr::read_csv(datafile)
```
