Reading Data from R

Pittsburgh Bridges Data Set

The data set is part of the UCI Machine Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges

The data Description is here:

http://archive.ics.uci.edu/ml/machine-learning-databases/bridges/bridges.names

The Data file is here:

http://archive.ics.uci.edu/ml/machine-learning-databases/bridges/bridges.data.version1

Read the description, and take a look at the data set. Some of the usual questions you should try to keep in mind:

- Is there a row for column names?
- Is there a column for row names?
- What is the character used as field separator?
- What is the character used for decimal point?
- Are there any missing values? If so, how are they codified?
- What is the data type of each variable (i.e. column)?

Reading the data with `read.table()`:

- Download a copy of the data to your computer (use `download.file()`) and save it in a file named `bridges.data`
- Create a vector of column names, to be used for the argument `col.names`
- Create a vector of column types,to be used for the argument `colClasses`
- Turn off conversion of strings as factors
- Use the function `read.table()` to read the data. Name it `bridges`.
- Once you've read the table, check its structure with `str()`

Now reread `bridges.data` with the function `read.csv()`

Low-level function `scan()`

- One of the low-level functions to read data values is `scan()`.
- The first argument is `file`
- The second argument is `what`
- If all the data values are of the same type, `what` usually takes the value of the data type (e.g. `"numeric"`, or `"character"`)

- If there is a mix of data types, then `what` should be a list indicating the data types for each field.

Let's use `scan()` to try to read `bridges.data`

NTSB Aviation Data

The National Transportation Safety Board (NTSB) has an aviation accident database containing information from 1962. The data set can be downloaded from the following url:

http://app.ntsb.gov/aviationquery/download.ashx?type=csv

The desription of the fields (data dictionary) can be found here:

http://www.ntsb.gov/_layouts/ntsb.aviation/AviationDownloadDataDictionary.aspx

I subset the lines corresponding to 2015, and saved the file in the github repo inside the `data/` folder under the name `aviation-2015.txt`:

https://raw.githubusercontent.com/ucb-stat243/stat243-fall-2016/master/data/aviation-2015.txt

If you take a look at the file, you will see that the fields are separated with a vertical bar surrounded by two spaces: `" | "`.

How would you import this table in R?

Airline Data

Consider the data file `airlineSubsample.csv` from the 2016 R bootcamp taught by Chris Paciorek.

https://raw.githubusercontent.com/berkeley-scf/r-bootcamp-2016/master/data/airlineSubsample.csv

Consider the following questions:

- What is the total number of flights (# of lines) in `airlineSubsample.csv`
- What is the number of flights in 2005?
- What is the total number of flights for each available year?
- Subset lines of 2005 flights in a separate csv file

Let's answer the previous questions using R, and then compare them with solutions using bash commands.