

Big Data: Data Analysis Boot Camp

Assumed File Structure

Chuck Cartledge, PhD

19 January 2018

Table of contents (1 of 1)

- 1 Intro.
- 2 Overall
- 3 Scripts
- 4 Data
- 5 Temporary things

What are we going to cover?

Most of the R scripts used in this boot camp assume a certain directory structure, and specific locations for scripts, data files, and images. To wit:

- 1 Scripts directory – where R scripts “live”
- 2 Data directory – where data usually comes from and goes to

If images are created, they are considered a type of data.

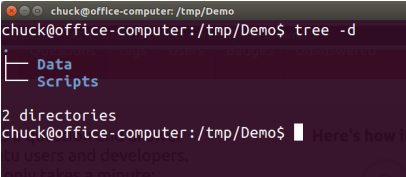


Relationship of directories.

There are two directories, relative to one another, and anywhere in the file system.

- Data – where data files are usually read from or written to
- Scripts – where R scripts are executed and “sourced” from

The file and directory names are Unix case sensitive and should be safe in a Windows environment.

A terminal window titled 'chuck@office-computer: /tmp/Demo' showing the command 'tree -d' and its output. The output shows a tree structure with two subdirectories: 'Data' and 'Scripts'. Below the tree, it says '2 directories' and the prompt 'chuck@office-computer: /tmp/Demo\$' is visible. There is also some faint text below the prompt that says 'Here's how to...' and 'for users and developers' and 'in a Windows environment'.

```
chuck@office-computer: /tmp/Demo
chuck@office-computer: /tmp/Demo$ tree -d
.
├── Data
└── Scripts

2 directories
chuck@office-computer: /tmp/Demo$
```

Where programs execute.

- “Base” R scripts may source() other script files. All files are assumed to live in the scripts directory.
- Access to data files within the R scripts is via the file.path() function.

| [ICO] | Name | Last modified | Size | Description |
|-------------|--|------------------|------|-------------|
| [PARENTDIR] | Parent Directory | - | - | - |
| [] | airTraffic.R | 2017-12-15 15:28 | 5.5K | |
| [] | anscombe.R | 2017-12-15 15:28 | 666 | |
| [] | cancerData.R | 2017-12-15 15:28 | 12K | |
| [] | chapter-01-ansleubing.R | 2017-12-15 15:28 | 6.5K | |
| [] | chapter-03-ansleubing.R | 2017-12-15 15:28 | 5.0K | |
| [] | chapter-04-crime-cluster.R | 2017-12-15 15:28 | 1.1K | |
| [] | chapter-04-life-cluster.R | 2017-12-15 15:28 | 3.3K | |
| [] | chapter-04-life-expectancy.R | 2017-12-15 15:28 | 1.0K | |
| [] | chapter-04.R | 2017-12-15 15:28 | 1.6K | |
| [] | chapter-05-life-expectancy.R | 2017-12-15 15:28 | 1.4K | |
| [] | chapter-05-mvsvin-voting.R | 2017-12-15 15:28 | 2.6K | |
| [] | chapter-05-trucks-help.R | 2017-12-15 15:28 | 339 | |
| [] | chapter-05-trucks.R | 2017-12-15 15:28 | 1.3K | |
| [] | chapter-06-mysql.R | 2017-12-15 15:28 | 6.5K | |

Apache Server at www.cs.ou.edu Port 80

All file accesses should be Operating System agnostic.

Sometimes, things need to be persistent.

- Data files live, and die in the data directory.
- The Data directory is one “up” from the Scripts directory.
- All accesses to the data directory are via the `file.path()` function.

| [ICO] | Name | Last modified | Size | Description |
|-------------|--|------------------|------|-------------|
| [PARENTDIR] | Parent Directory | - | - | - |
| [] | 133241921_T_T10RD_MARKET_US_CARRIER_ONLY.zip | 2017-12-15 15:31 | 475K | |
| [] | 247805599_T_T10RD_MARKET_US_CARRIER_ONLY.zip | 2017-12-15 15:31 | 1.6M | |
| [] | 9781792169352_code.zip | 2017-12-15 15:31 | 772K | |
| [] | 9781786466457_Code.zip | 2017-12-15 15:32 | 78M | |
| [TXT] | DizemZ.txt | 2017-12-15 15:32 | 335 | |
| [TXT] | StrandsPackt.csv | 2017-12-15 15:32 | 17M | |
| [] | corpus.dat | 2017-12-15 15:32 | 3.1M | |
| [IMG] | heart-outline.png | 2017-12-15 15:32 | 18K | |
| [] | msl-est2016-01 | 2017-12-15 15:32 | 17K | |
| [] | msl-est2016-01.xlsx | 2017-12-15 15:32 | 17K | |
| [] | shdls_1.0.8.zip | 2017-12-15 15:32 | 67K | |
| [] | nomos-hof-judict_base64 | 2017-12-15 15:32 | 202K | |
| [] | swiss_votes.dat | 2017-12-15 15:32 | 1.7K | |

Apache Server at www.cs.odu.edu Port 80

All file accesses should be Operating System agnostic.

Where temporary things live.

- There are times when data does not need to be persistent.
- Generally these data are stored wherever `tempfile()` or `tempdir()` put them.
- Be aware that data stored, either directly or indirectly using these functions may be removed when the R session completes.

All file accesses should be Operating System agnostic.