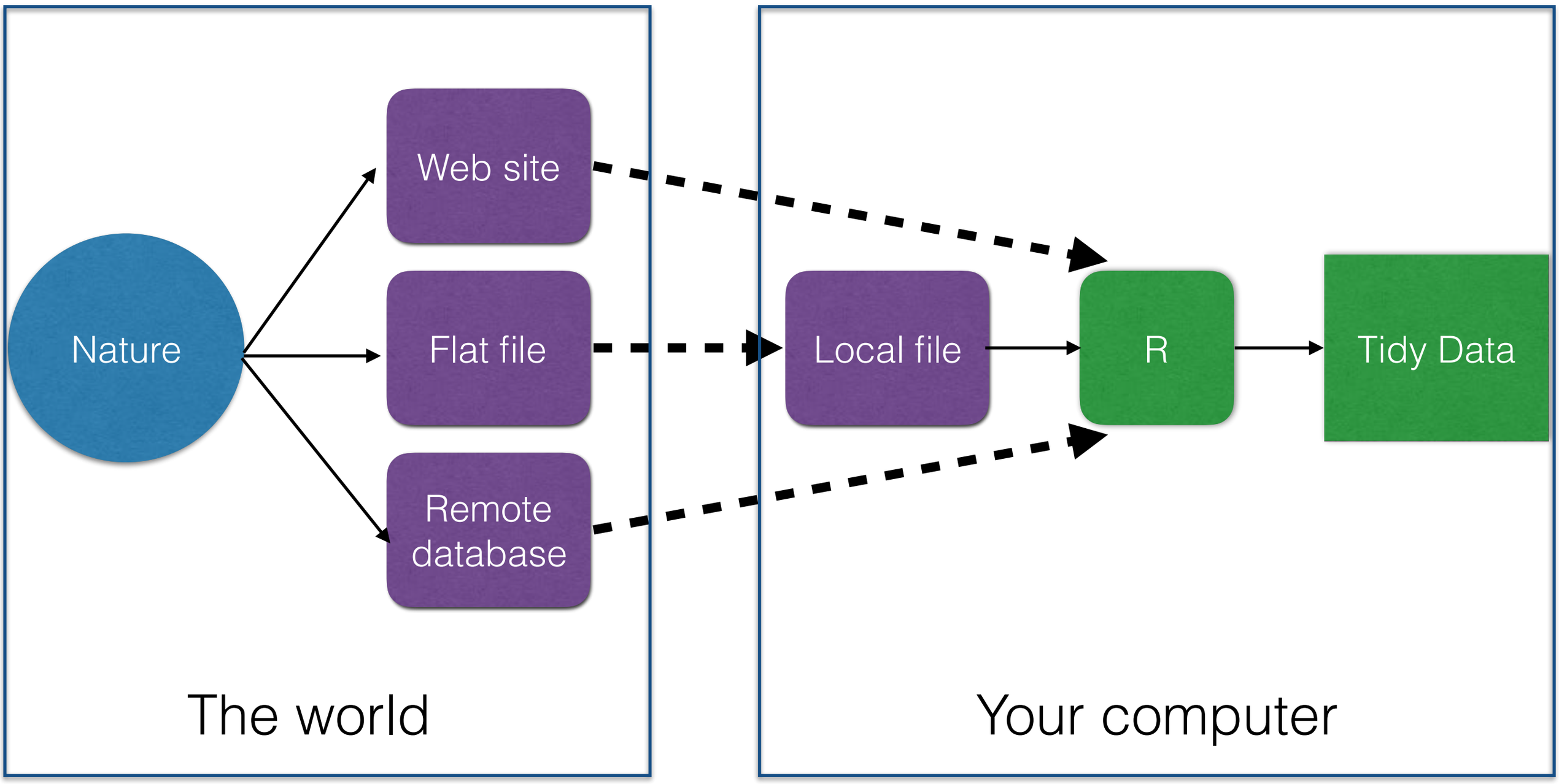
Getting and Cleaning Data

Biostatistics 140.776

Getting and Cleaning Data

- Getting data: APIs and web scraping
- Cleaning data: Tidy data
- Transforming data: Regular expressions

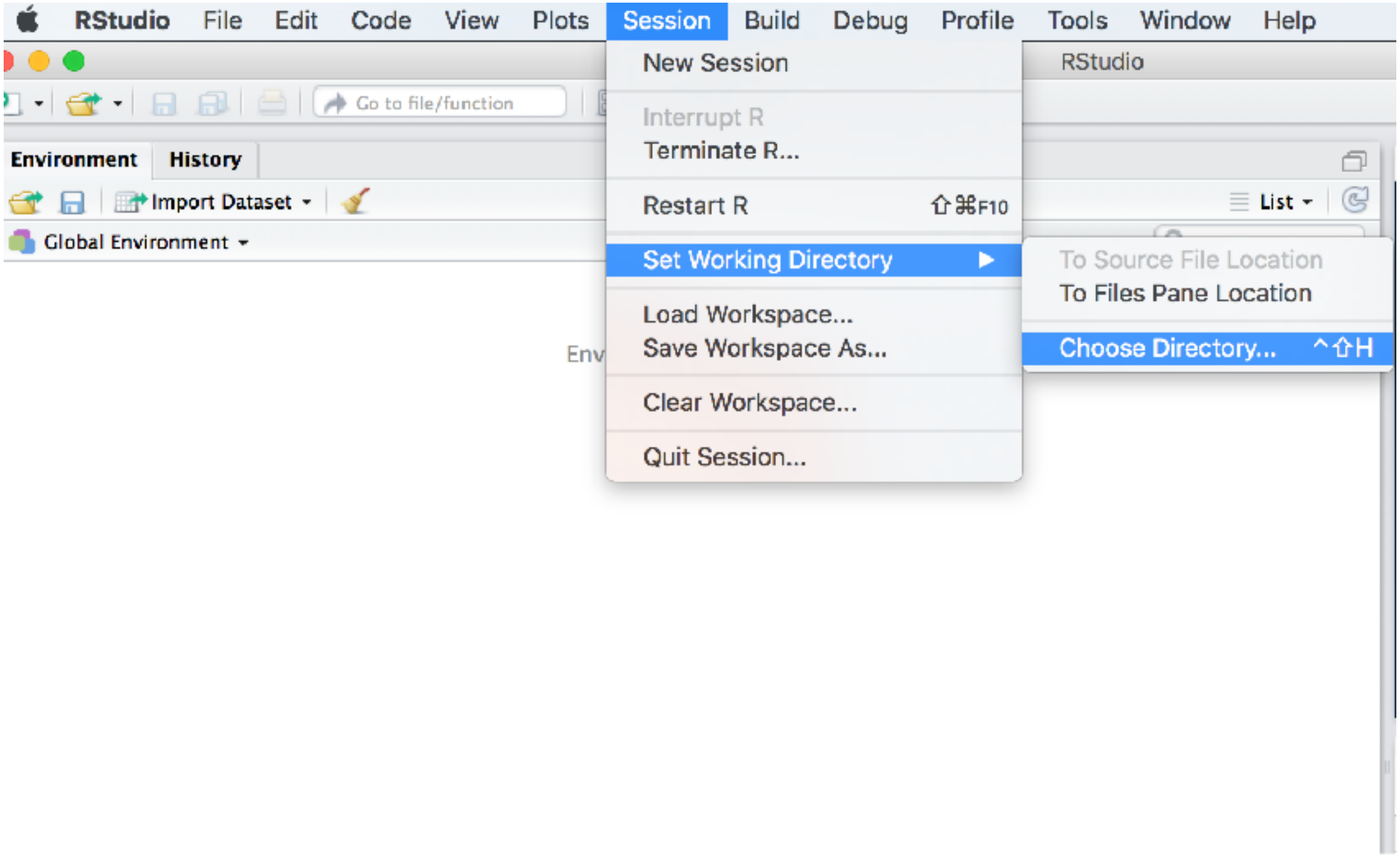
Getting Data



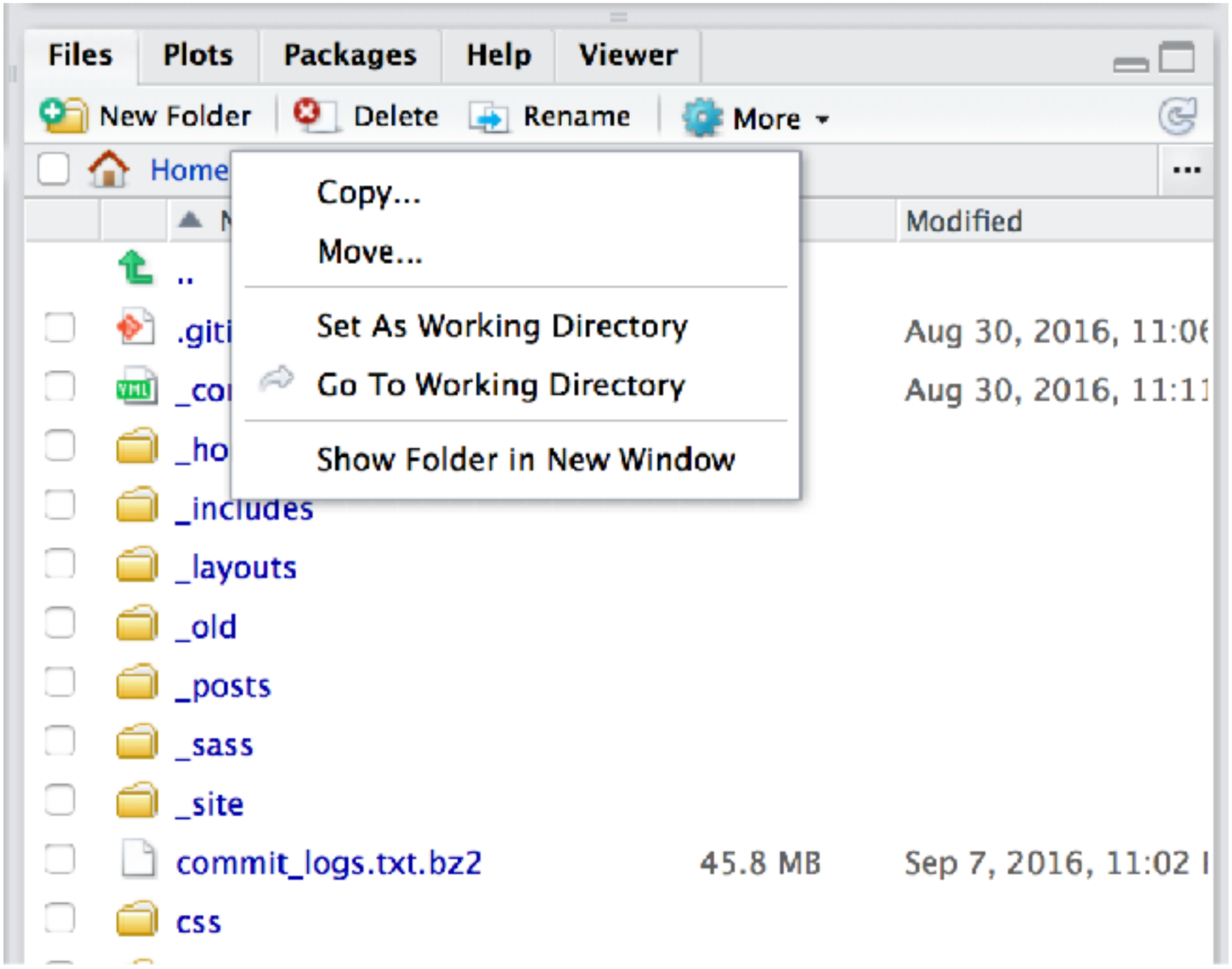
Where Am I?

- The **working directory** is the place where R will look when reading or writing files
- `getwd()` returns the current working directory
- `setwd()` sets the working directory
- RStudio has menu options

Working Directory



Working Directory



Working Directory

- When you download files/data, move them to your working directory

- Upon opening R, **set your working directory** to match whatever project you are working on

- Much simpler if you use RStudio Projects

- Check with `getwd()`

The readr Package

- The functions in the **readr** package can be used to read tabular data from text files

- `read_csv`, `read_table` are the core functions

- All functions are very fast and *can read directly from compressed files*

- You should try to never decompress files that are compressed (not much to be gained)

Flat Files

Particulates

Year	PM2.5 FRM/FEM Mass (88101)	PM2.5 non FRM/FEM Mass (88502)	PM10 Mass (81102)	PM2.5 Speciation
2016	daily_88101_2016.zip 77,865 Rows 932 KB As of 2016-06-17	daily_88502_2016.zip 53,685 Rows 649 KB As of 2016-06-17	daily_81102_2016.zip 30,980 Rows 251 KB As of 2016-06-17	daily_SPEC_2016.zip 92,441 Rows 824 KB As of 2016-06-17
2015	daily_88101_2015.zip 390,090 Rows 4,213 KB As of 2016-06-17	daily_88502_2015.zip 311,901 Rows 3,454 KB As of 2016-06-17	daily_81102_2015.zip 167,829 Rows 1,216 KB As of 2016-06-17	daily_SPEC_2015.zip 1,690,870 Rows 13,061 KB As of 2016-06-17
2014	daily_88101_2014.zip 369,476 Rows 3,979 KB As of 2016-06-17	daily_88502_2014.zip 333,364 Rows 3,675 KB As of 2016-06-17	daily_81102_2014.zip 167,191 Rows 1,221 KB As of 2016-06-17	daily_SPEC_2014.zip 2,108,467 Rows 16,717 KB As of 2016-06-17

http://aqsdrl.epa.gov/aqsweb/aqstmp/airdata/download_files.html

Flat Files

Protocol

URL

```
download.file("http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/
daily_88101_2016.zip", “PM2.5_2016.zip")
```

Local file name

```
PM2.5_trying URL 'http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/
daily_88101_2016.zip'
Content type 'application/zip' length 954069 bytes (931 KB)
==================================================
downloaded 931 KB
```

Flat Files

```
## Look inside the zip file (without decompressing it)
unzip("PM2.5_2016.zip", list = TRUE)

                    Name   Length                 Date
1 daily_88101_2016.csv 27720545 2016-07-07 19:35:00
```

```
## Read the file (again without decompressing)
library(readr)
d <- read_csv(unz("PM2.5_2016.zip", "daily_88101_2016.csv"))
```

Create connection to zip archive file

Name of zip file

File name within zip archive

Web Sites

Liquor Licenses
(No description provided)

Find in this Dataset

Manage | More Views | Filter | Visualize | Export | Discuss | Embed | About

	LicenseClass	SubClass	LicenseNumber	LicenseDate	LicenseEndDate	LicenseYear	LicenseFee	CertificateNumber	LicenseStatus	LicenseeFirstN
2	LBD7	BWL	243	05/01/2015	04/30/2016	2015	$1,320.00	656	Renewed	SALLY
3	LBD7	BWL	341	05/01/2015	04/30/2016	2015	$1,320.00	924	Renewed	JASON
4	LBD7	BWL	81	05/01/2015	04/30/2016	2015	$1,320.00	224	Renewed	WAYNE
5	LC	BWL	51	05/01/2015	04/30/2016	2015	$550.00	1077	Renewed	EARLE A.
6	LC	BWL	51	05/01/2015	04/30/2016	2015	$550.00	1077	Renewed	JAMES A.
7	LC	BWL	51	05/01/2015	04/30/2016	2015	$550.00	1077	Renewed	JOHN C.
8	LBD7	BWL	304	05/01/2015	04/30/2016	2015	$1,320.00	827	Renewed	FRED A.
9	LBD7	BWL	304	05/01/2015	04/30/2016	2015	$1,320.00	827	Renewed	SHIRLEY O.
10	LBD7	BWL	242	05/01/2015	04/30/2016	2015	$1,320.00	655	Renewed	ADRIENNE M.
11	LBD7	BWL	242	05/01/2015	04/30/2016	2015	$1,320.00	655	Renewed	TONYA M.
12	AE	AE	16	07/01/2015	06/30/2016	2015	$1,000.00	1236	Renewed	ADRIENNE M.
13	AE	AE	16	07/01/2015	06/30/2016	2015	$1,000.00	1236	Renewed	TONYA M.
14	WD	BW	17	05/01/2015	04/30/2016	2015	$165.00	429	Renewed	ANNE
15	WD	BW	17	05/01/2015	04/30/2016	2015	$165.00	429	Renewed	MILTON J.
16	LA	BWL	88	05/01/2015	04/30/2016	2015	$858.00	356	Renewed	JONG WOONG
17	LA	BWL	88	05/01/2015	04/30/2016	2015	$858.00	356	Renewed	YONG JIN
18	LBD7	BWL	140	05/01/2015	04/30/2016	2015	$1,320.00	336	Renewed	MILTON W.

Look for this

https://data.baltimorecity.gov/City-Services/Liquor-Licenses/xv8d-bwgi

Web Sites

Liquor Licenses
(No description provided)

Find in this Dataset

Manage | More Views | Filter | Visualize | Export | Discuss | Embed | About

	LicenseClass	SubClass	LicenseNumber	LicenseDate	LicenseEndDate	LicerseYear	LicenseFee	Certific
2	LBD7	BWL	243	05/01/2015	04/30/2016	2015	$1,320.00	
3	LBD7	BWL	341	05/01/2015	04/30/2016	2015	$1,320.00	
4	LBD7	BWL	81	05/01/2015	04/30/2016	2015	$1,320.00	
5	LC	BWL	51	05/01/2015	04/30/2016	2015	$550.00	
6	LC	BWL	51	05/01/2015	04/30/2016	2015	$550.00	
7	LC	BWL	51	05/01/2015	04/30/2016	2015	$550.00	
8	LBD7	BWL	304	05/01/2015	04/30/2016	2015	$1,320.00	
9	LBD7	BWL	304	05/01/2015	04/30/2016	2015	$1,320.00	
10	LBD7	BWL	242	05/01/2015	04/30/2016	2015	$1,320.00	
11	LBD7	BWL	242	05/01/2015	04/30/2016	2015	$1,320.00	
12	AE	AE	16	07/01/2015	06/30/2016	2015	$1,000.00	
13	AE	AE	16	07/01/2015	06/30/2016	2015	$1,000.00	
14	WD	BW	17	05/01/2015	04/30/2016	2015	$165.00	
15	WD	BW	17	05/01/2015	04/30/2016	2015	$165.00	
16	LA	BWL	88	05/01/2015	04/30/2016	2015	$858.00	
17	LA	BWL	88	05/01/2015	04/30/2016	2015	$858.00	
18	LBD7	BWL	140	05/01/2015	04/30/2016	2015	$1,320.00	

Export

SODA API

OData

Print

Download

Download a copy of this dataset in a static format

Download As

CSV

CSV for Excel

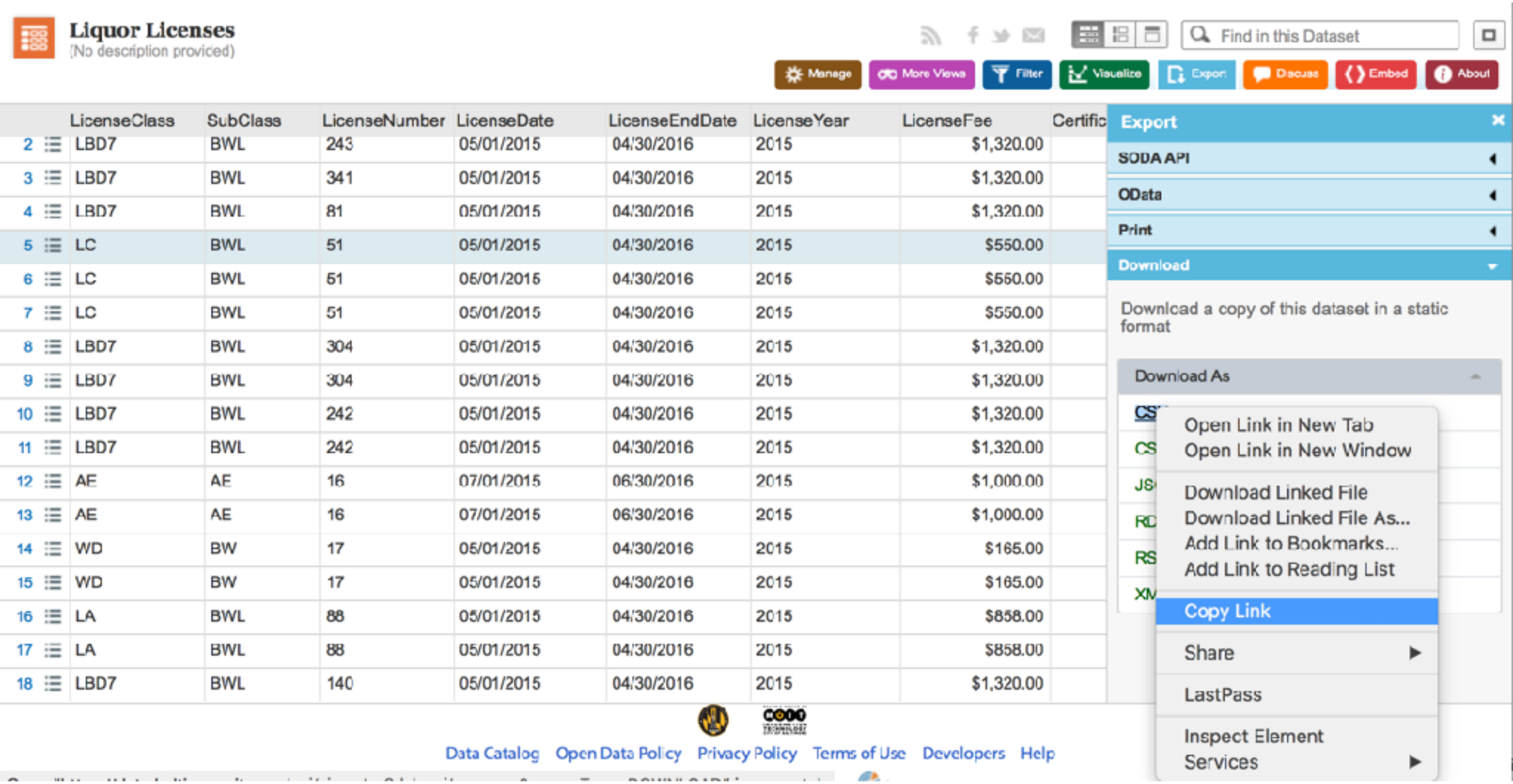
JSON

RDF

RSS

XML

Web Sites


Liquor Licenses
(No description provided)
Find in this Dataset
Manage
More Views
Filter
Visualize
Export
Discuss
Embed
About
LicenseClass SubClass LicenseNumber LicenseDate LicenseEndDate LicenseYear LicenseFee Certific
2 LBD7 BWL 243 05/01/2015 04/30/2016 2015 $1,320.00
3 LBD7 BWL 341 05/01/2015 04/30/2016 2015 $1,320.00
4 LBD7 BWL 81 05/01/2015 04/30/2016 2015 $1,320.00
5 LC BWL 51 05/01/2015 04/30/2016 2015 $550.00
6 LC BWL 51 05/01/2015 04/30/2016 2015 $550.00
7 LC BWL 51 05/01/2015 04/30/2016 2015 $550.00
8 LBD7 BWL 304 05/01/2015 04/30/2016 2015 $1,320.00
9 LBD7 BWL 304 05/01/2015 04/30/2016 2015 $1,320.00
10 LBD7 BWL 242 05/01/2015 04/30/2016 2015 $1,320.00
11 LBD7 BWL 242 05/01/2015 04/30/2016 2015 $1,320.00
12 AE AE 16 07/01/2015 06/30/2016 2015 $1,000.00
13 AE AE 16 07/01/2015 06/30/2016 2015 $1,000.00
14 WD BW 17 05/01/2015 04/30/2016 2015 $165.00
15 WD BW 17 05/01/2015 04/30/2016 2015 $165.00
16 LA BWL 88 05/01/2015 04/30/2016 2015 $858.00
17 LA BWL 88 05/01/2015 04/30/2016 2015 $858.00
18 LBD7 BWL 140 05/01/2015 04/30/2016 2015 $1,320.00
Export
SODA API
OData
Print
Download
Download a copy of this dataset in a static format
Download As
Open Link in New Tab
Open Link in New Window
Download Linked File
Download Linked File As...
Add Link to Bookmarks...
Add Link to Reading List
Copy Link
Share
LastPass
Inspect Element
Services
Data Catalog
Open Data Policy
Privacy Policy
Terms of Use
Developers
Help

Read a File from Web

```
library(readr)

## Read directly off the web site
liquor <- read_csv("https://data.baltimorecity.gov/api/views/xv8d-bwgi/
rows.csv?accessType=DOWNLOAD")

## Note the date/time data were downloaded
Sys.time()
```

Reading Remote Data

- Functions in the **readr** package (`read_csv`, `read_table`, etc.) can all read directly off the web

- Whether to retain static file or read dynamically off the web depends on application

- Truly “dynamic” applications (continuously updated) should probably read directly off web

- When you need a snapshot (e.g. for reproducibility), better to download a static file and read locally

Reading Other Files

- The **readxl** package can be used to read Microsoft Excel files (via the `read_excel` function)

- The **xlsx** package can also be used (via read.xlsx function) but requires Java and sometimes doesn't install

- The **haven** package can read files from other stats packages (`read_dta`, `read_sas`, `read_spss`)

- Also, the **foreign** package can read from a few other systems

There's a Package for That

- > 10,000 packages on CRAN
- **googlesheets**: Reading Google Sheets table data
- **jpeg**, **png**: Reading bitmap image data
- **rgdal**, **raster**, **shapefiles**: GIS data
- **tuneR**, **seewave**: Music/Sound data
- **RMySQL**, **SQLite**: Relational databases

When in Doubt…

The `read_lines` function just reads lines of text

```
library(readr)

## Read lines from a text file (commit logs)
logs <- read_lines("commit_logs_nomerge.txt")

## See the first few lines of text
head(logs)
[1] "commit 7f6ef08e80191712a5eb0d75c42931466e7bbe73"
[2] “Author: XXXXXXXXXX"
[3] "Date:   Wed Oct 1 16:55:12 2014 -0400"
[4] ""
[5] "    date changes to pages/tickets"
[6] ""
```

Raw Text Data

- Useful for taking a peek at data because no processing is done on read-in
- Use text processing tools (regular expressions) to convert to tidy data
- Lots of back and forth between examining the **structure** of text data and manipulating that structure
- Structured data, like XML, has its own packages for reading and processing

Goal: Tidy Data

- Allows for easy manipulation, transformation, and summary of data

- Required format for most modeling functions (`lm`, `glm`, etc.)

- Makes plotting data in various layouts/formats easier (i.e. with **ggplot2** package)

- Easier to reason about the data

- Everything is easier! 🦄🌈

Tidy Data

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

https://www.jstatsoft.org/article/view/v059i10

Tidy Data

	Rural Male	Rural Female	Urban Male	Urban Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

VADeaths dataset

Tidy Data

```
      age  urban gender death_rate
   <fctr> <fctr> <fctr>      <dbl>
1   50-54  Rural   Male       11.7
2   55-59  Rural   Male       18.1
3   60-64  Rural   Male       26.9
4   65-69  Rural   Male       41.0
5   70-74  Rural   Male       66.0
6   50-54  Rural Female        8.7
7   55-59  Rural Female       11.7
```

Tidyverse

- A collection of packages that allows one to easily work with or create tidy data

- The **tidyr** package has tools for manipulating “wide” and “long” format data

- Many other packages in tidyverse depend on tidy data format: **dplyr**, **ggplot2**

Summary

- Identify data locations and tools for reading
- Download files to **working directory**
- Read data directly off the web (sometimes)
- Get data into tidy format