was chatting with some cyber-mates at a recent event and the topic of cyber attacks on the U.S. power-grid came up (as it often does these days). The conversation was brief, but the topic made its way into active memory and resurfaced when I saw today’s [Data Is Plural](https://tinyletter.com/data-is-plural/letters/data-is-plural-2018-09-12-edition) newsletter which noted that *“Utility companies are required to report major power outages and other “electric disturbance events” to the Department of Energy within a business day (or, depending on the type of event, sooner) of the incident. The federal agency then aggregates the reports annual summary datasets.”* (follow the links to the newsletter to get the URLs for the site since Jeremy deserves your ).

Many of us data nerds use the Data Is Plural newsletters as fodder for class assignments, blog posts or personal “data katas”. This time, I was after cyber attack data.

When you head to the annual reports URL, you’re greeted with a SharePoint-driven HTML table:

[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/html-table/)

So, our options are PDF or XLS (and I mean .xls, too, they’re not modern .xlsx files). We’ll opt for the latter and cache them locally before working on them. One “gotcha” is that the hrefs look like this: https://www.oe.netl.doe.gov/download.aspx?type=OE417XLS&ID=78 — i.e. no filenames. *But*, the filenames come along for the ride when an HTTP GET or HEAD request is issued in a [content-disposition response header](https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Content-Disposition). We’ll use this metadata instead of siphoning off the year from the first column of the table:

library(rvest)

library(readxl)

library(tidyverse)

doe <- read\_html("https://www.oe.netl.doe.gov/OE417\_annual\_summary.aspx")

dir.create("~/Data/doe-cache-dir", showWarnings = FALSE)

html\_nodes(doe, xpath=".//a[contains(., 'XLS')]") %>%

html\_attr("href") %>%

{ .pb <<- progress\_estimated(length(.)) ; . } %>% # we likely don't rly need progress bars tho

walk(~{

.pb$tick()$print()

dl\_url <- sprintf("https://www.oe.netl.doe.gov/%s", .x)

res <- HEAD(dl\_url) # so we can get the filename

stop\_for\_status(res) # halt on network errors

fil <- str\_replace(

string = res$headers['content-disposition'],

pattern = "attachment; filename=",

replacement = "~/Data/doe-cache-dir/"

)

if (!file.exists(fil)) { # this pattern allows us to issue a lightweight HTTP HEAD request, then cache and refresh w/o wasting server/our bandwidth/cpu

res <- GET(dl\_url, httr::write\_disk(fil))

stop\_for\_status(res)

Sys.sleep(5) # be kind to the server(s) but only if we're downloading data files since HEAD requests don't really tax services

}

})

Let's do a quick check for the likelihood of uniformity. Some of these files go back to 2002 and I suspect they're more geared for "printing" (the PDF counterparts were a clue) than programmatic processing:

# check to see if the files are all the same (spoiler alert: they're not)

list.files("~/Data/doe-cache-dir", "xls", full.names=TRUE) %>%

map\_df(~list(

fil = basename(.x),

ncols = read\_xls(.x, col\_names=FALSE, col\_types="text") %>% ncol()

)) -> cols\_profile

cols\_profile

## # A tibble: 17 x 2

## fil ncols

##

## 1 2002\_Annual\_Summary.xls 8

## 2 2003\_Annual\_Summary.xls 8

## 3 2004\_Annual\_Summary.xls 8

## 4 2005\_Annual\_Summary.xls 8

## 5 2006\_Annual\_Summary.xls 8

## 6 2007\_Annual\_Summary.xls 8

## 7 2008\_Annual\_Summary.xls 8

## 8 2009\_Annual\_Summary.xls 8

## 9 2010\_Annual\_Summary.xls 8

## 10 2011\_Annual\_Summary.xls 9

## 11 2012\_Annual\_Summary.xls 9

## 12 2013\_Annual\_Summary.xls 9

## 13 2014\_Annual\_Summary.xls 9

## 14 2015\_Annual\_Summary.xls 11

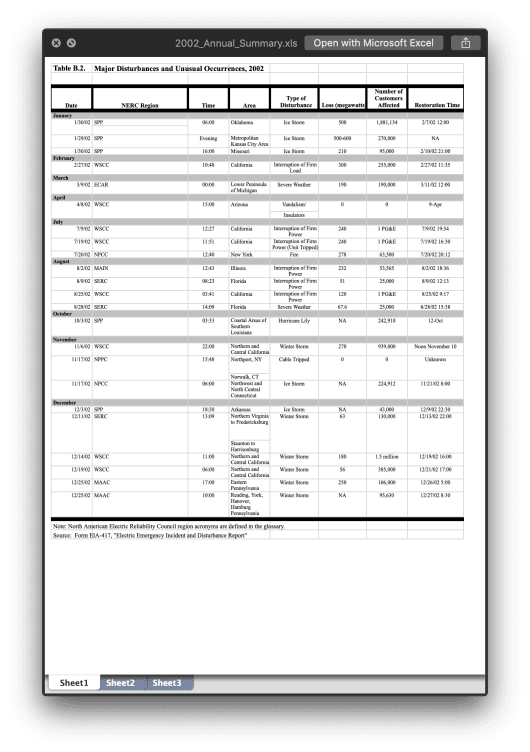
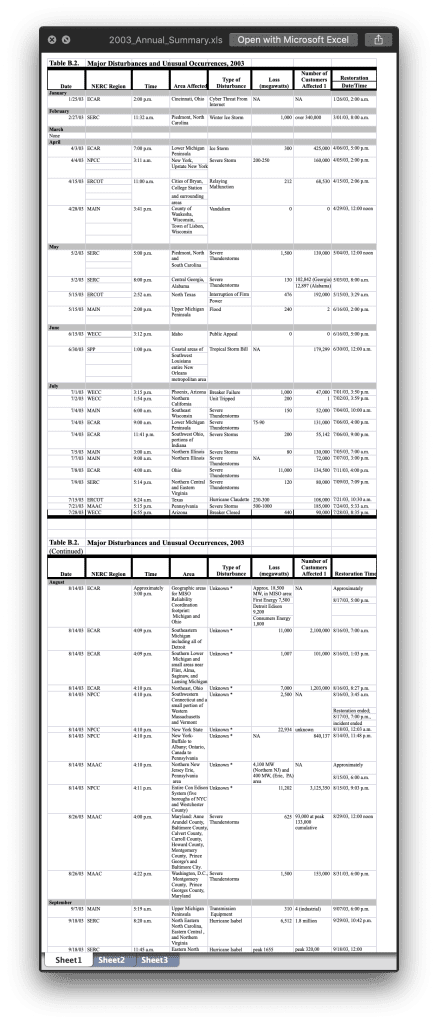
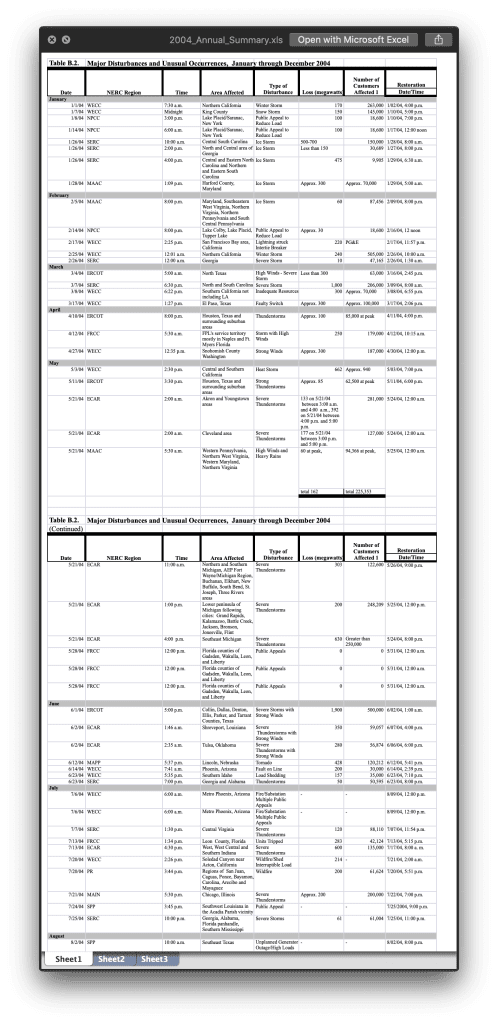
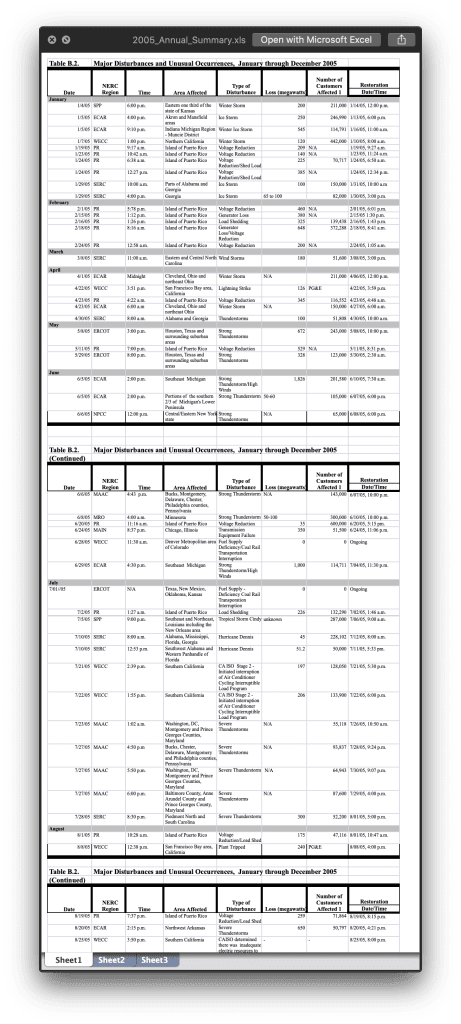
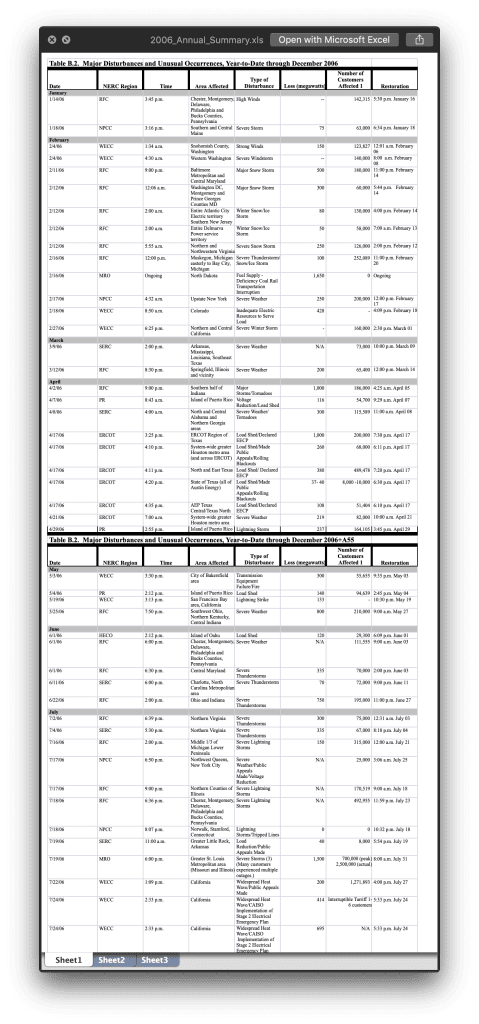
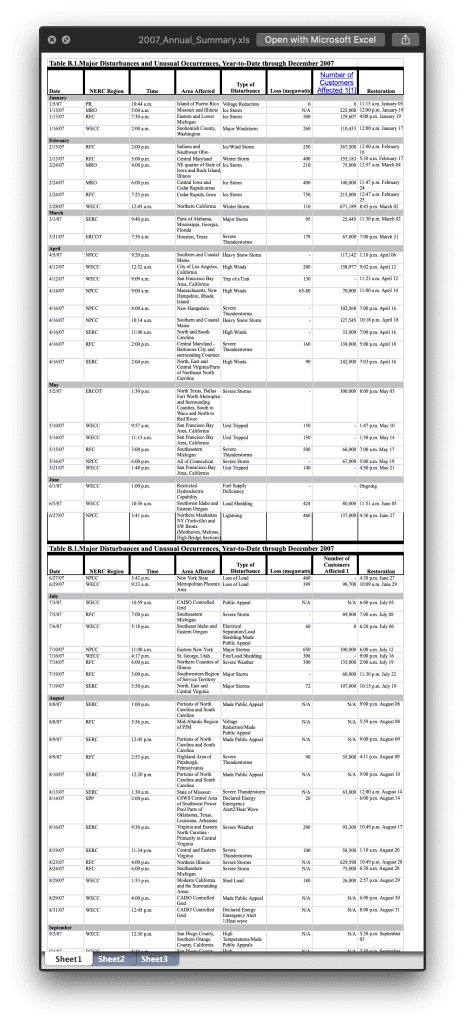
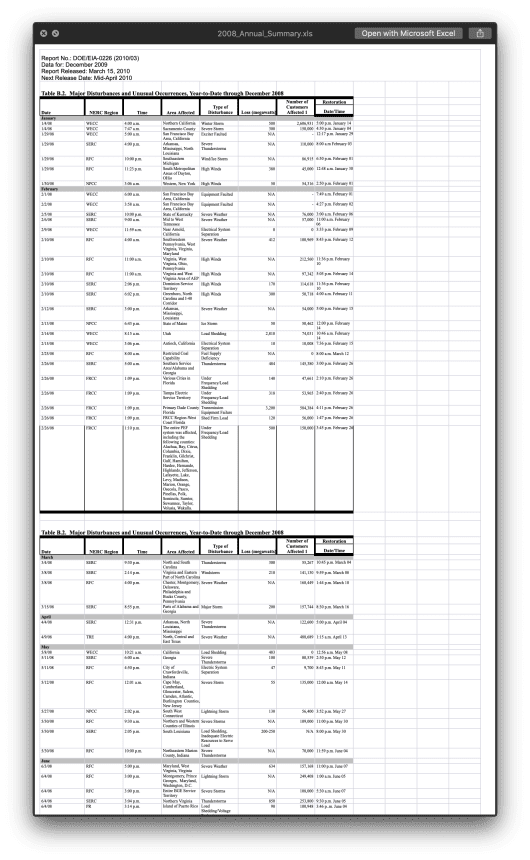
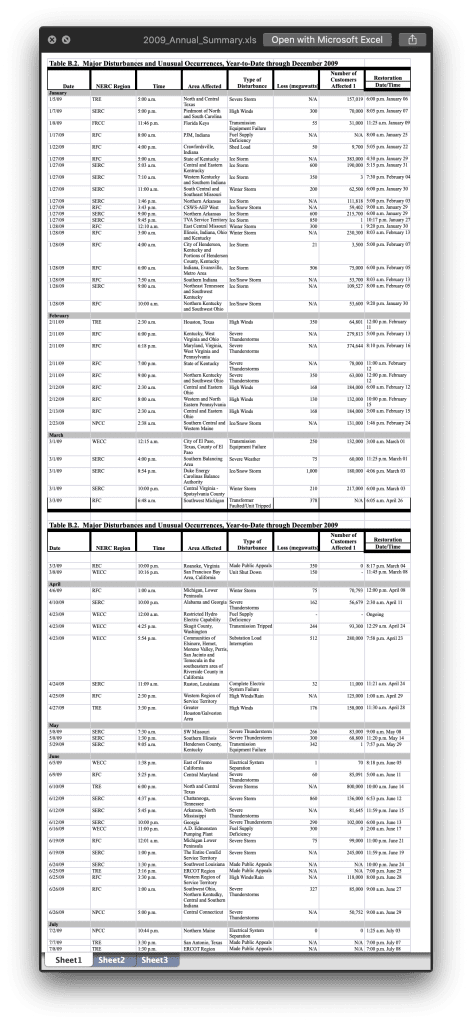
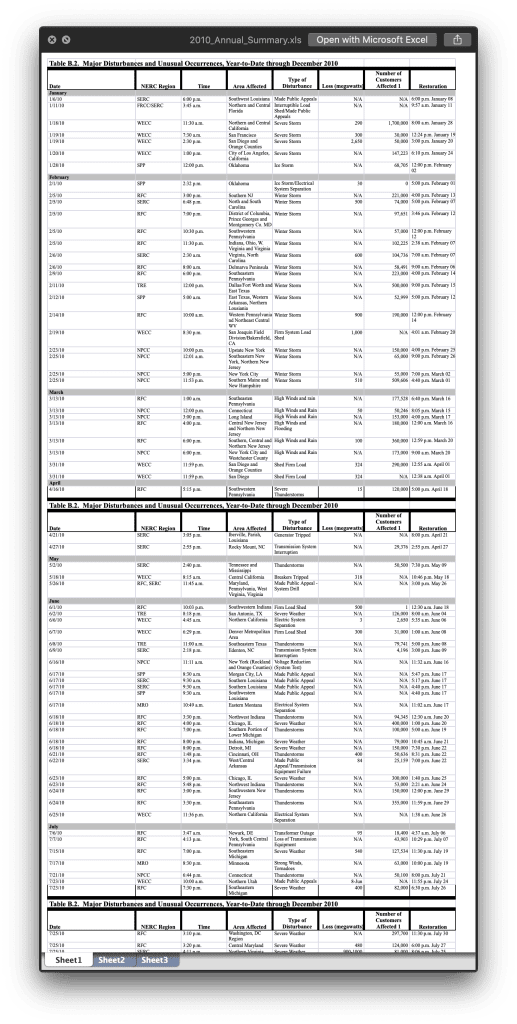
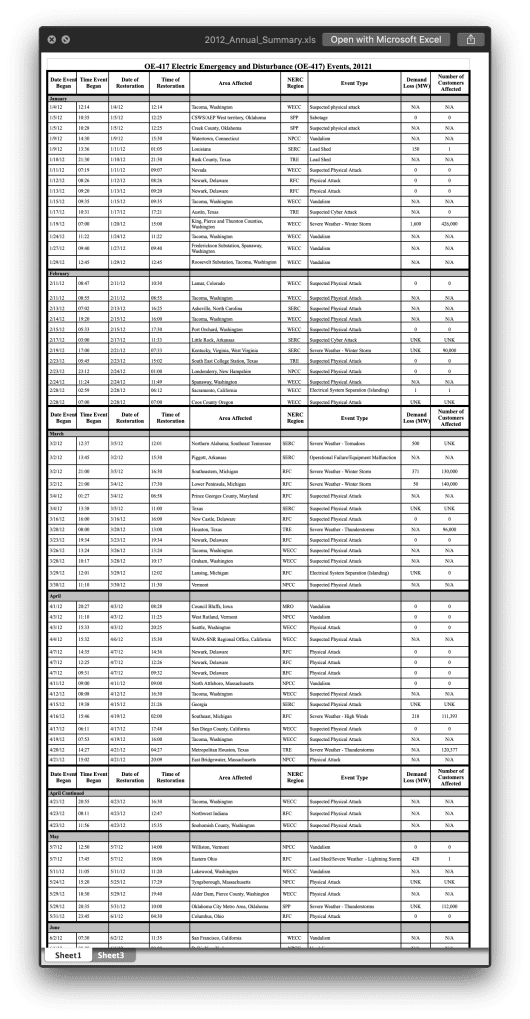
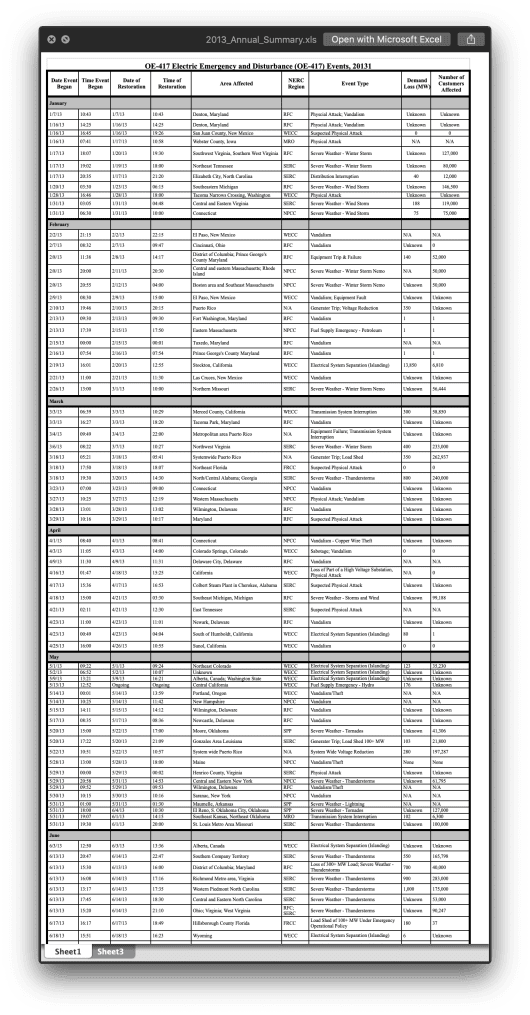
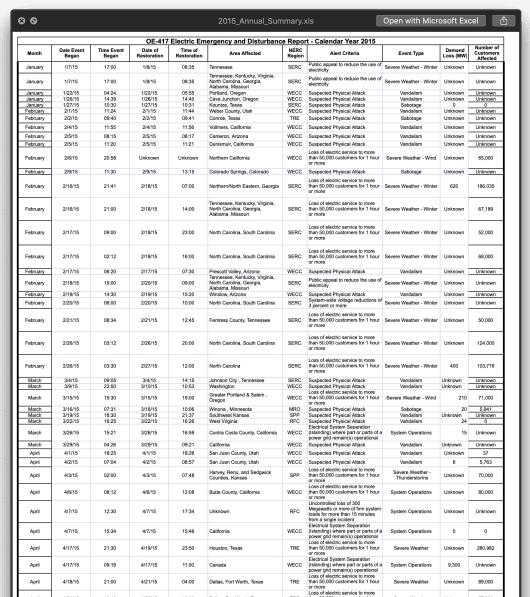
## 15 2016\_Annual\_Summary.xls 11

## 16 2017\_Annual\_Summary.xls 11

## 17 2018\_Annual\_Summary.xls 11

O\_o

At this point, I paused and wanted to see what was going on in the minds of the DoE staffers charged with releasing this data.

[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2002/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2003/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2004/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2005/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2006/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2007/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2008/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2009/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2010/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2011/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2012/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2013/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2014/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2015/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2016/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2017/)  
[](https://rud.is/b/2018/09/12/the-evolution-of-data-literacy-at-the-u-s-department-of-energy-finding-power-grid-cyber-attacks-in-a-data-haystack/attachment/2018/)

(You can grab the macOS Quick Look preview snaps of all of those [here](https://rud.is/dl/doe-evolution.zip).)

From 2002 to 2010 the Excel documents are clearly designed for print as the target, complete with month breaklines and repeated (+ heavily formatted) headers. They even left other tabs around (inconsistently).

Things got a *little* better between 2011 and 2014, but we still have month breaks and occasional, repeated headers (someone likely learned how to generate headers-per-page in Excel in 2011 then the administration changed hands and new staffers came in and fubar'd 2012 a bit before going back to the slightly better format).

Prior to 2015, the print-as-target trumped programmatic access. Interestingly enough, this is roughly when "data science" was on the upswing (in a major way):

﻿

Starting with 2015 we have a "month" column, more uniformity for specifying dates & times and more care given to other value fields, so kudos to the then and current staffers who made our data-machinating lives easier.

This really is, I believe, a byproduct of modern "data literacy". Folks in charge of gathering and publishing data are realizing there are multiple ways others want/need to consume the data. The original purpose for this data was to hand a report to someone after regulations were put in place to mandate notifications. I'm willing to bet nobody did anything with this data for a few years. Staffers either learned to wield Excel better or new staffers came in with this new knowledge. Finally, recent years clearly show that the staffers realize that folks are as (or more) likely to programmatically consume this information as they are reading a a long list of events (). More work is needed (and an API or CSV/JSON output would be super cool) but it's great to see data literacy alive and well in the halls of the U.S. gov.

Said modern format changes do not really help us work with the complete data set and the more recent files have issues all their own, including inconsistency in the way the date/time columns are represented in Excel cells.

By golly, we're *still* going to try to read all these files in and work with them (for at least the purpose I originally set out on). We'll have to deal with the differences in columns and come up with a way to remove non-data rows. I also kinda at least want dates as dates. Here's my stab at an initial clean-up (there's lots more work to do, though):

map2(cols\_profile$fil, cols\_profile$ncols, ~{

if (.y == 8) { # handle 8 cols

suppressWarnings(read\_xls(

path = sprintf("~/Data/doe-cache-dir/%s", .x),

col\_names = c("date\_began", "region", "time", "area", "event\_type", "loss", "customers\_affected", "date\_restored"),

col\_types = c("date", "text", "text", "text", "text", "text", "text", "date")

)) %>%

filter(!is.na(date\_began)) %>%

mutate(date\_began = as.Date(date\_began))

} else if (.y == 9) { # handle 9 cols

suppressWarnings(read\_xls(

path = sprintf("~/Data/doe-cache-dir/%s", .x),

col\_names = c("date\_began", "time\_began", "date\_restored", "time\_restored", "area", "region", "event\_type", "loss", "customers\_affected"),

col\_types = c("date", "guess", "date", "guess", "text", "text", "text", "text", "text")

)) %>%

filter(!is.na(date\_began)) %>%

mutate(date\_began = as.Date(date\_began))

} else if (.y == 11) { # handle 11 cols

# note that the date columns aren't uniform in the Excel spreadsheets even in these more data-literate files :-(

suppressWarnings(read\_xls(

path = sprintf("~/Data/doe-cache-dir/%s", .x),

col\_names = c("month", "date\_began", "time\_began", "date\_restored", "time\_restored", "area", "region", "alert\_criteria", "event\_type", "loss", "customers\_affected"),

col\_types = c("text", "text", "guess", "text", "guess", "text", "text", "text", "text", "text", "text")

)) %>%

mutate(

date\_began = case\_when(

str\_detect(date\_began, "/") ~ suppressWarnings(as.Date(date\_began, format="%m/%d/%Y")),

str\_detect(date\_began, "^[[:digit:]]+$") ~ suppressWarnings(as.Date(as.integer(date\_began), origin = "1899-12-30")),

TRUE ~ suppressWarnings(as.Date(NA))

)

) %>%

mutate(

date\_restored = case\_when(

str\_detect(date\_restored, "/") ~ suppressWarnings(as.Date(date\_restored, format="%m/%d/%Y")),

str\_detect(date\_restored, "^[[:digit:]]+$") ~ suppressWarnings(as.Date(as.integer(date\_restored), origin = "1899-12-30")),

TRUE ~ suppressWarnings(as.Date(NA))

)

) %>%

filter(!is.na(date\_began))

}

}) -> reports

reports[[1]]

## # A tibble: 23 x 8

## date\_began region time area event\_type loss customers\_affec… date\_restored

##

## 1 2002-01-30 SPP 0.25 Oklahoma Ice Storm 500 1881134 2002-02-07 12:00:00

## 2 2002-01-29 SPP Evening Metropol… Ice Storm 500-… 270000 NA

## 3 2002-01-30 SPP 0.66666666666666663 Missouri Ice Storm 210 95000 2002-02-10 21:00:00

## 4 2002-02-27 WSCC 0.45000000000000001 Californ… Interruption o… 300 255000 2002-02-27 11:35:00

## 5 2002-03-09 ECAR 0 Lower Pe… Severe Weather 190 190000 2002-03-11 12:00:00

## 6 2002-04-08 WSCC 0.625 Arizona Vandalism/ 0 0 2002-04-09 00:00:00

## 7 2002-07-09 WSCC 0.51875000000000004 Californ… Interruption o… 240 1 PG&E 2002-07-09 19:54:00

## 8 2002-07-19 WSCC 0.49375000000000002 Californ… Interruption o… 240 1 PG&E 2002-07-19 16:30:00

## 9 2002-07-20 NPCC 0.52777777777777779 New York Fire 278 63500 2002-07-20 20:12:00

## 10 2002-08-02 MAIN 0.52986111111111112 Illinois Interruption o… 232 53565 2002-08-02 18:36:00

## # ... with 13 more rows

reports[[10]]

## # A tibble: 307 x 9

## date\_began time\_began date\_restored time\_restored area region event\_type loss customers\_affec…

##

## 1 2011-01-11 0.96388888… 2011-01-11 00:00:00 0.96388888888… Athens… NPCC Electrica… 0 0

## 2 2011-01-12 0.25 2011-01-12 00:00:00 0.58333333333… Massac… NPCC Winter St… N/A 80000

## 3 2011-01-13 0.30624999… 2011-01-13 00:00:00 0.34236111111… North … FRCC Firm Syst… 150 20900

## 4 2011-01-18 0.58333333… 2011-01-18 00:00:00 0.58333333333… Whitma… NPCC Vandalism 0 0

## 5 2011-01-23 0.29166666… 2011-01-23 00:00:00 0.54166666666… Frankl… WECC Vandalism 0 0

## 6 2011-01-24 0.55555555… 2011-01-24 00:00:00 0.5625 Newman… WECC Suspiciou… 0 0

## 7 2011-01-25 0.14097222… 2011-01-25 00:00:00 0.45833333333… Newark… RFC Vandalism 0 0

## 8 2011-01-26 0.39236111… 2011-01-27 00:00:00 0.70833333333… Carson… WECC Suspected… 0 0

## 9 2011-01-26 0.39791666… 2011-01-27 00:00:00 0.62708333333… Michig… RFC Vandalism 0 0

## 10 2011-01-26 0.70833333… 2011-01-31 00:00:00 0.33333333333… Montgo… RFC Winter St… N/A 210000

## # ... with 297 more rows

reports[[17]]

## # A tibble: 120 x 11

## month date\_began time\_began date\_restored time\_restored area region alert\_criteria event\_type loss

##

## 1 Janua… 2018-01-01 0.7645833… 2018-01-02 0.7576388888… Tenne… SERC Public appeal to … Severe We… Unkn…

## 2 Janua… 2018-01-01 0.7381944… NA Unknown Texas: TRE Public appeal to … Severe We… Unkn…

## 3 Janua… 2018-01-01 0.9006944… 2018-01-02 0.4375 Tenne… SERC Public appeal to … System Op… Unkn…

## 4 Janua… 2018-01-02 0.4166666… 2018-02-12 0.3333333333… New Y… NPCC Fuel supply emerg… Fuel Supp… 675

## 5 Janua… 2018-01-02 0.3125 NA Unknown South… SERC Public appeal to … Severe We… 0

## 6 Janua… 2018-01-02 0.28125 2018-01-02 0.375 North… SERC System-wide volta… Severe We… 14998

## 7 Janua… 2018-01-04 0.0756944… 2018-01-04 0.0895833333… Texas… TRE Physical attack t… Actual Ph… Unkn…

## 8 Janua… 2018-01-12 0.5472222… 2018-01-12 0.6201388888… Michi… RF Cyber event that … System Op… 41

## 9 Janua… 2018-01-15 0.1805555… 2018-01-18 0.2416666666… Texas: TRE Public appeal to … Severe We… Unkn…

## 10 Janua… 2018-01-16 0.625 2018-01-18 0.5416666666… Tenne… SERC Public appeal to … Severe We… Unkn…

## # ... with 110 more rows, and 1 more variable: customers\_affected

If you'd've handled the above differently it'd be if you could drop a note in the comments (for both my benefit and that of any other readers who have kindly made it this far into this tome).

At this point, I really just want to *finally* see if there are any "cyber" events in the data set and when/where they were. To do that, let's whittle down the columns a bit and make one data frame out of all the reports:

map\_df(reports, ~{

select(.x, date\_began, region, area, event\_type, customers\_affected, date\_restored) %>%

mutate(date\_restored = as.Date(date\_restored)) %>%

mutate(

customers\_affected = suppressWarnings(

str\_replace\_all(customers\_affected, "\\-.\*$|[[:punct:]]+|[[:alpha:]]+", "") %>%

as.numeric()

)

) %>%

mutate(date\_restored = as.Date(ifelse(is.na(date\_restored), date\_began, date\_restored), origin = "1970-01-01"))

}) -> events

events

## # A tibble: 2,243 x 6

## date\_began region area event\_type customers\_affect… date\_restored

##

## 1 2002-01-30 SPP Oklahoma Ice Storm 1881134 2002-02-07

## 2 2002-01-29 SPP Metropolitan Kansas City Area Ice Storm 270000 2002-01-29

## 3 2002-01-30 SPP Missouri Ice Storm 95000 2002-02-10

## 4 2002-02-27 WSCC California Interruption of Firm Load 255000 2002-02-27

## 5 2002-03-09 ECAR Lower Peninsula of Michigan Severe Weather 190000 2002-03-11

## 6 2002-04-08 WSCC Arizona Vandalism/ 0 2002-04-09

## 7 2002-07-09 WSCC California Interruption of Firm Power 1 2002-07-09

## 8 2002-07-19 WSCC California Interruption of Firm Powe… 1 2002-07-19

## 9 2002-07-20 NPCC New York Fire 63500 2002-07-20

## 10 2002-08-02 MAIN Illinois Interruption of Firm Power 53565 2002-08-02

## # ... with 2,233 more rows

Now we're cookin' with gas!

Let's do a quick check to make sure things look OK:

count(events, event\_type, sort=TRUE)

## # A tibble: 390 x 2

## event\_type n

##

## 1 Severe Weather 369

## 2 Vandalism 216

## 3 Severe Weather - Thunderstorms 97

## 4 Suspected Physical Attack 87

## 5 System Operations 74

## 6 Severe Thunderstorms 70

## 7 Winter Storm 51

## 8 Ice Storm 42

## 9 Physical Attack - Vandalism 40

## 10 High Winds 33

## # ... with 380 more rows

Those events+quantities seem to make sense. Now, for my ultimate goal:

filter(events, grepl("cyber|hack", event\_type, ignore.case=TRUE)) # yep, grepl() is still in muscle memory

## # A tibble: 19 x 6

## date\_began region area event\_type customers\_affec… date\_restored

##

## 1 2003-01-25 ECAR Cincinnati, Ohio Cyber Threat From Int… NA 2003-01-25

## 2 2011-02-03 RFC Bowie, Maryland Suspected Cyber Attack 0 2011-02-03

## 3 2011-02-17 WECC Roseville, California Suspected Cyber Attack 0 2011-02-23

## 4 2011-03-14 RFC Baltimore, Maryland Suspected Cyber Attack NA 2011-03-14

## 5 2011-04-03 SERC Unknown Suspected Cyber Attack 0 2011-04-05

## 6 2011-07-08 RFC PJM Corporate Office, Pennsylvania Suspected Cyber Attack NA 2011-07-11

## 7 2011-12-21 WECC Boise, Idaho Suspected Cyber Attack 0 2011-12-21

## 8 2012-01-17 TRE Austin, Texas Suspected Cyber Attack 0 2012-01-17

## 9 2012-02-17 SERC Little Rock, Arkansas Suspected Cyber Attack NA 2012-02-17

## 10 2012-11-15 MRO Iowa; Michigan Suspected Cyber Attack NA 2012-11-15

## 11 2013-06-21 MRO Michigan, Iowa Suspected Cyber Attack NA 2013-10-30

## 12 2013-10-16 SERC Roxboro Plant, North Carolina Cyber Event with Pote… 0 2013-10-16

## 13 2014-03-20 NPCC New York Suspected Cyber Attack NA 2014-03-20

## 14 2014-10-21 MRO Carmel, Indiana Suspected Cyber Attack NA 2014-10-21

## 15 2014-12-30 NPCC New Hampshire, Massachusetts, Mai… Suspected Cyber Attack NA 2014-12-31

## 16 2016-02-07 NPCC New York: Orange County Cyber Attack NA 2016-02-07

## 17 2016-04-12 WECC Washington: Pend Oreille County Cyber Attack 0 2016-04-12

## 18 2016-11-09 WECC California: Stanislaus County, Sa… Cyber Attack 0 2016-11-09

## 19 2016-12-13 WECC California: Riverside County; Cyber Event 0 2016-12-13



**FIN**

There's a great deal of work left out of this power-outage data cleanup work:

* Turn outage start/end info into POSIXct objects
* Normalize area (make it a data\_frame column with state and municipality so it can be unnested nicely)
* Normalize event\_type since many of the phrases used are equivalent and some have more than one categorization
* Normalize loss somehow and do a better job with customers\_affected (I did not double-check my work and I think there are errors in that column, now, but I didn't need it for my goal).