DATA GETS PERSONAL

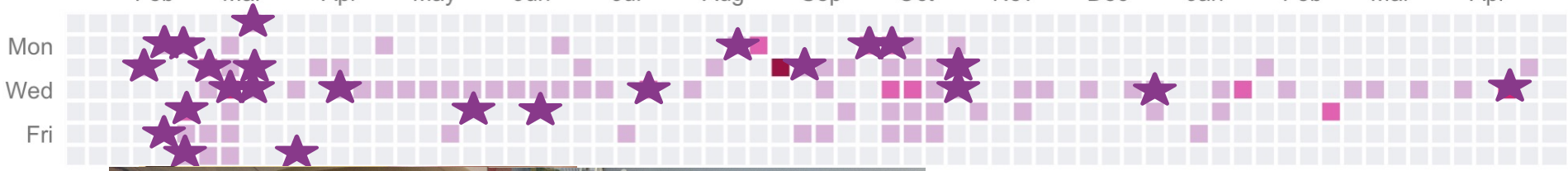
Louisa H. Smith

PhD candidate in epidemiology
Harvard TH Chan School of Public Health

January 29, 2019

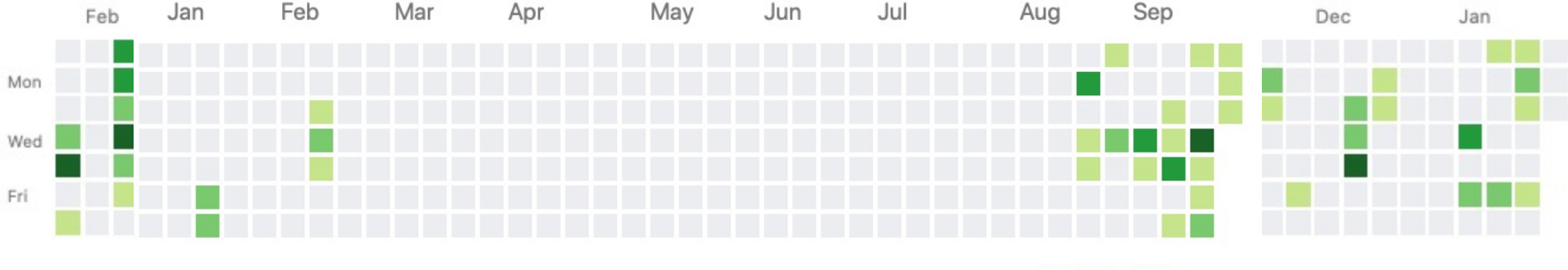
Tonight's goal

Tell a
(data science?
human interest?)
story with R.


Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
Mon
Wed
Fri


Trastuzumab
600
mL/hr
BID

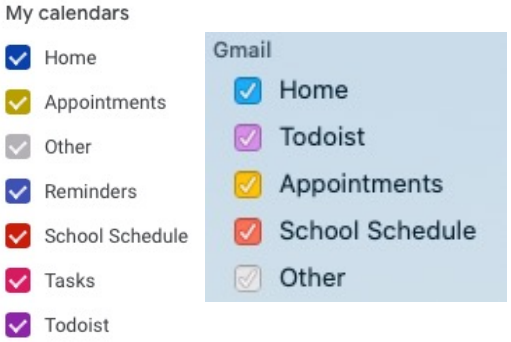
Inspiration
(what it actually looked like during that time)



Process

Export calendar as .ics

I keep all my medical appointments as a separate calendar
Exporting is easy with Google Calendars, iCal, I'm sure others



Convert .ics to .csv

I used an online tool: http://www.indigoblue.eu/ics2csv/

Clean data!

I got some help getting it in the format I want for plotting from the source code of this blog post: https://www.garrickadenbuie.com/blog/greatest-twitter-scheme/

Full explanation here:
https://www.louisahsmith.com/post/github-style-calendar-heatmap/


Kim Kardashian West
@KimKardashian
Like butter. #Butter350's #Yeezy
64.8K 10:11 AM - Aug 6, 2018
8,805 people are talking about this

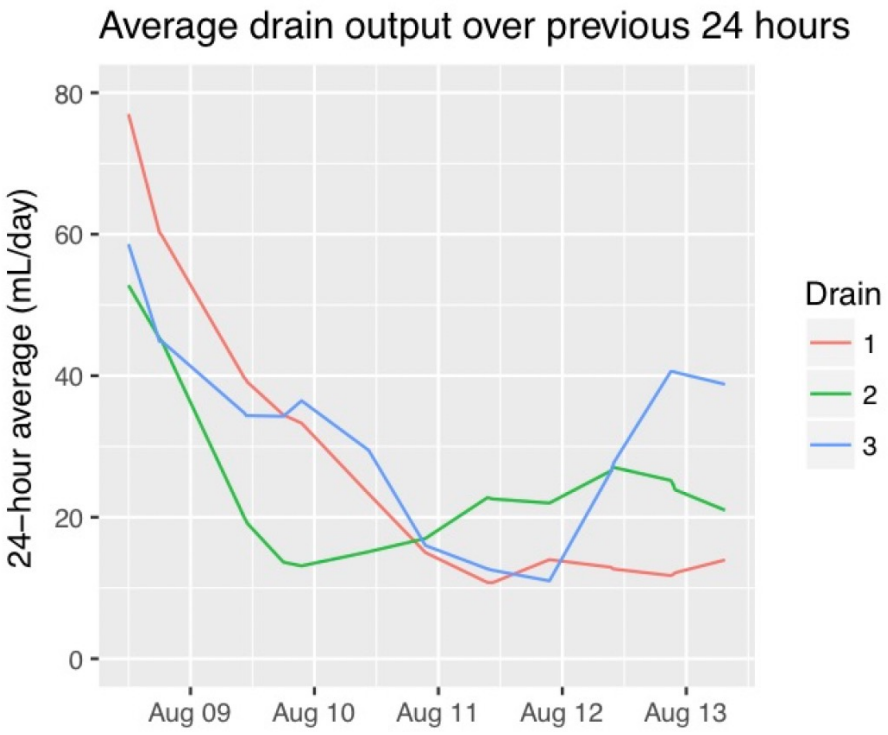
What they expected…

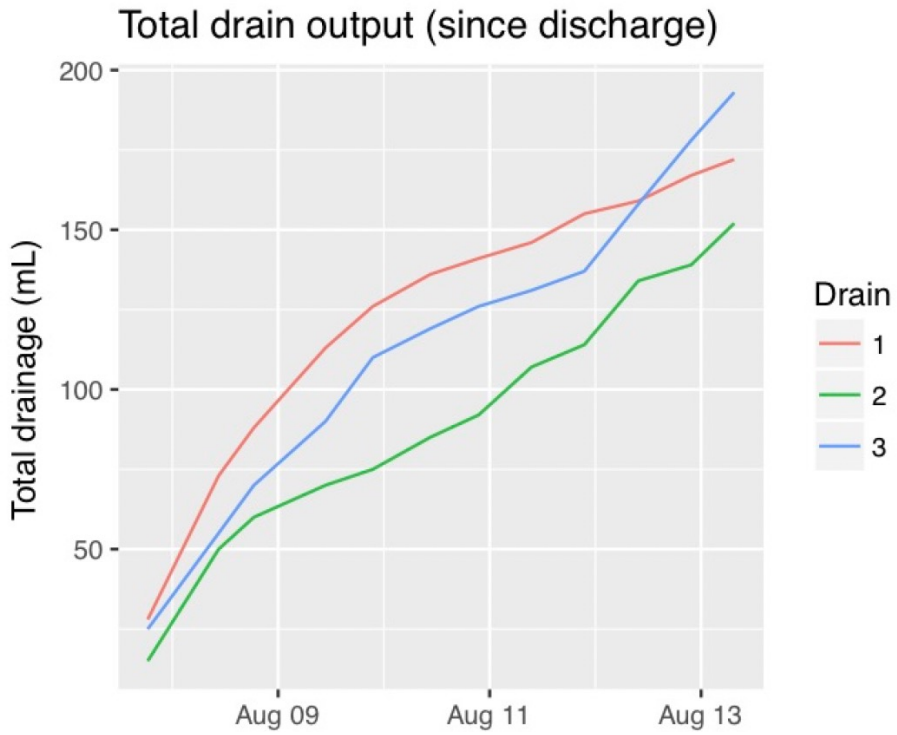
DATE	TIME	DRAIN #1	DRAIN #2	DRAIN #3	DRAIN #4
Location (ie: Breast, Abdomen, etc)					
8/7	6:25	28	15	25	

What they got…

```
drains ← tibble::tribble(
        ~date,      ~time, ~one, ~two, ~three,
  "08/07/2018",  "6:25 pm",   28,   15,     25,
  "08/08/2018", "10:30 am",   45,   35,     30,
  "08/08/2018",  "6:30 pm",   15,   10,     15,
  "08/09/2018", "10:45 am",   25,   10,     20,
  "08/09/2018",  "9:30 pm",   13,    5,     20,
  "08/10/2018", "10:30 am",   10,   10,      9,
  "08/10/2018",  "9:30 pm",    5,    7,      7,
  "08/11/2018",  "9:30 am",    5,   15,      5,
  "08/11/2018",  "9:30 pm",    9,    7,      6,
  "08/12/2018",  "9:45 am",    4,   20,     21,
  "08/12/2018",  "9:45 pm",    8,    5,     20,
  "08/13/2018",  "7:30 am",    5,   13,     15,
  "08/13/2018",  "9:45 pm",   NA,    4,     20,
  "08/14/2018", "11:10 am",   NA,   11,     20,
  "08/14/2018",  "9:00 pm",   NA,    4,     24,
  "08/15/2018",  "9:45 pm",   NA,   11,     36,
  "08/16/2018",  "9:15 pm",   NA,   11,     27,
  "08/17/2018",  "9:30 pm",   NA,   10,     13,
  "08/18/2018", "10:15 pm",   NA,   10,     12,
  "08/19/2018",  "9:45 pm",   NA,    7,     12,
  "08/20/2018", "10:10 pm",   NA,    6,     11,
  "08/22/2018", "10:30 pm",   NA,   16,     19,
  "08/23/2018", "10:00 pm",   NA,   18,     20,
  "08/24/2018", "10:45 pm",   NA,   13,      9,
  "08/25/2018",  "1:00 pm",   NA,   40,     15,
  "08/25/2018", "10:10 pm",   NA,   14,      6,
  "08/26/2018",  "9:30 pm",   NA,   10,     13,
  "08/27/2018", "11:00 pm",   NA,    7,     14,
  "08/28/2018", "10:30 pm",   NA,   NA,     20,
  "08/29/2018", "11:30 pm",   NA,   NA,     16,
  "08/30/2018", "11:00 pm",   NA,   NA,     16
  )
```

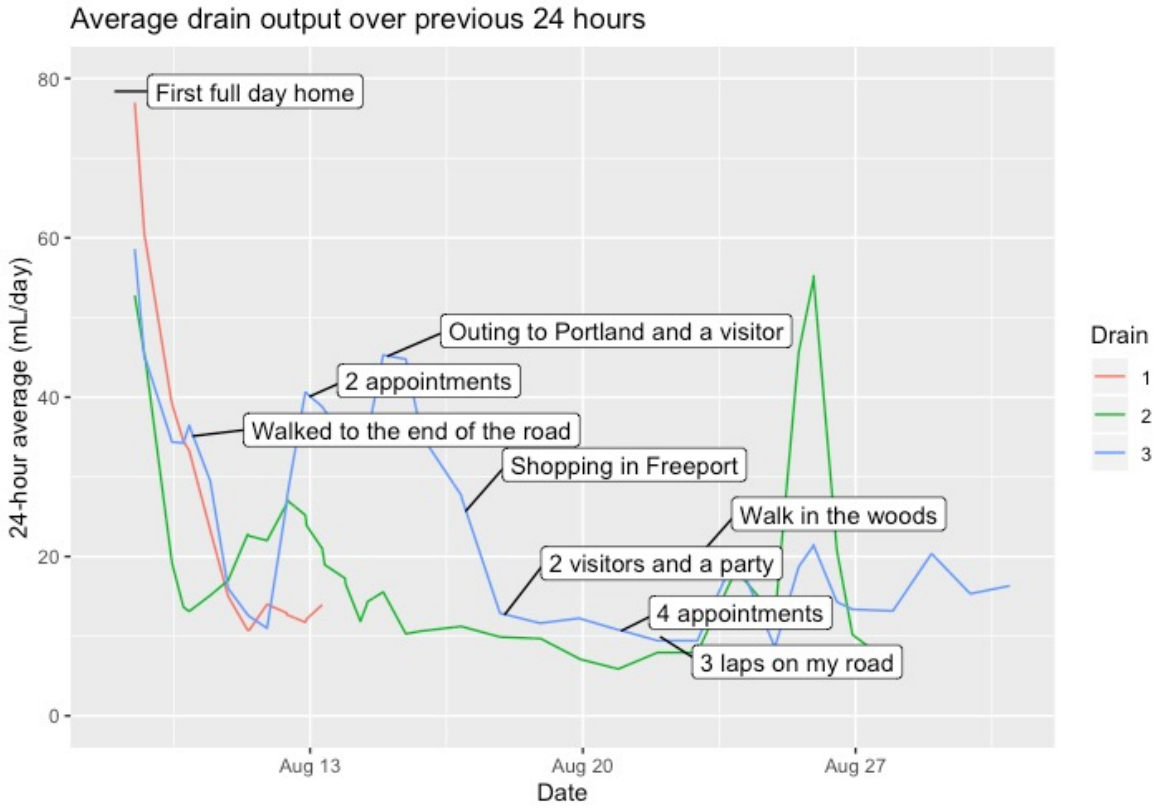
Removal is based on 24-hour output





Tried to see what correlated with high output

As you can imagine, drains are really uncomfortable….



Process

Collect data

`tibble::tribble()` is my go-to for on-the-spot data collection:

```
drains <- tibble::tribble(
         ~date,      ~time, ~one, ~two, ~three,
  "08/07/2018",  "6:25 pm",   28,   15,     25,
  "08/08/2018", "10:30 am",   45,   35,     30,
```

Kept notes on my phone and would move it to R whenever I had the chance

Use `datapasta` to keep nice and neat! (more later)

Clean data

Between `tidyr` and `lubridate`, easy creation of dates:

```
unite(date_time, c(date, time), sep = " ") %>%
mutate(date_time = mdy_hm(date_time))
```

`RcppRoll` for calculating rolling averages

`ggrepel` for adding labels to ggplots

Inside my patient portal…

COMPLETE BLOOD COUNT (BLOOD)

DATE	WBC 4.0-10.0 K/uL	RBC 3.9-5.2 m/uL	Hgb 11.2-15.7 g/dL	Hct 34-45 %	MCV 82-98 fL	MCH 26-32 pg	MCHC 32-37 g/dL	RDW 10.5-15.5 %	RDWSD 35.1-46.3 fL
04/04/18 9:30A (34)	5.0	3.77*	11.1*	32.5*	86	29.4	34.2	12.5	36.4
			(34) TY						
03/21/18 10:20A (36)	8.7	4.14	12.4	35.6	86	30.0	34.8	11.9	36.5
			(36) TY						
03/07/18 8:06A (38)	5.7	4.55	13.3	39.6	87	29.2	33.6	12.2	38.9
			(38) TY						
02/15/18 4:45P	10.0	4.95	14.7	43.1	87	29.7	34.1	11.9	38.0

DIFFERENTIAL (BLOOD)

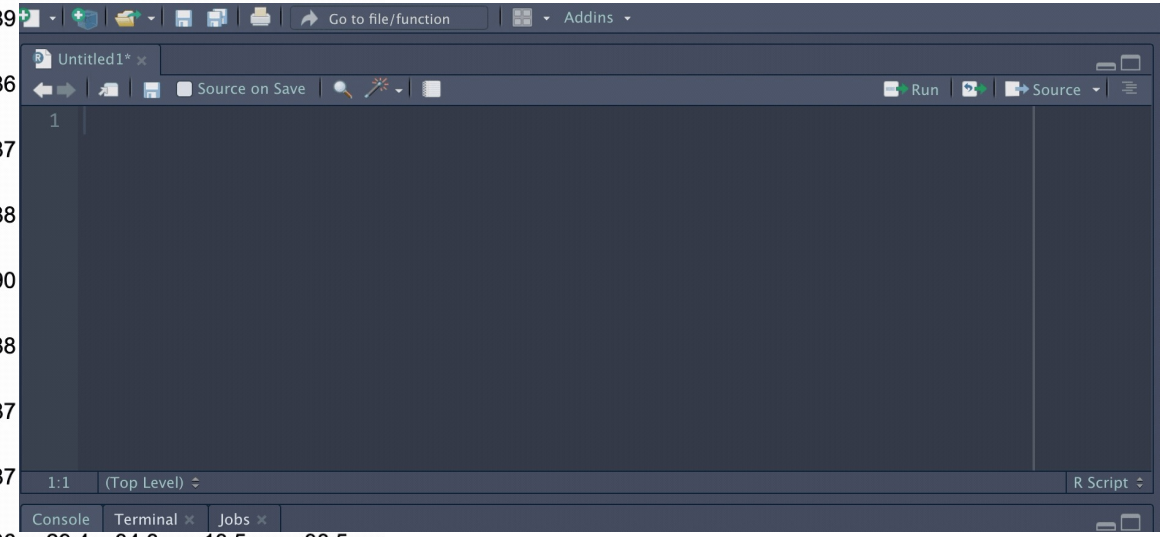
DATE	Neuts 34-71 %	Bands 0-5 %	Lymphs 19-53 %	Monos 5-13 %	Eos 1-7 %	Baso 0-1 %	Atyps 0-0 %	Metas 0-0 %	Myelos 0-0 %	Promyel 0-0 %	Young 0-0 %	Blasts 0-0 %	Hyperse 0-0 %	NRBC 0-0 %	Plasma 0-0 %	Histio %	LUC %	Im Gran 0-.6 %	Other 0-0 %	AbsNeut 1.6-6.1 K/uL	AbsLymp 1.2-3.7 K/uL	AbsMono .2-.8 K/uL	AbsEos .04-.54 K/uL	AbsBaso .01-.08 K/uL
07/18/19 3:05P	76.0*		17.9*	4.7*	0.5*	0.5												0.4		4.19	0.99*	0.26	0.03*	0.03
																		Includes Metas, Myelos, and Pros.						
12/19/18 9:45A (40)	78.2*		13.4*	6.8	0.5*	0.8												0.3		3.10	0.53*	0.27	0.02*	0.03
													(40) TY											
09/26/18 9:50A (42)																				2.18				
													(42) TY											
09/05/18 9:00A (44)																				1.94				
													(44) TY											
08/13/18 11:43A (46)																				3.05				
													(46) TY											
07/25/18 10:50A (48)																				5.68				
													(48) TY											

◄ ►

To RStudio…

COMPLETE BLOOD COUNT (BLOOD)

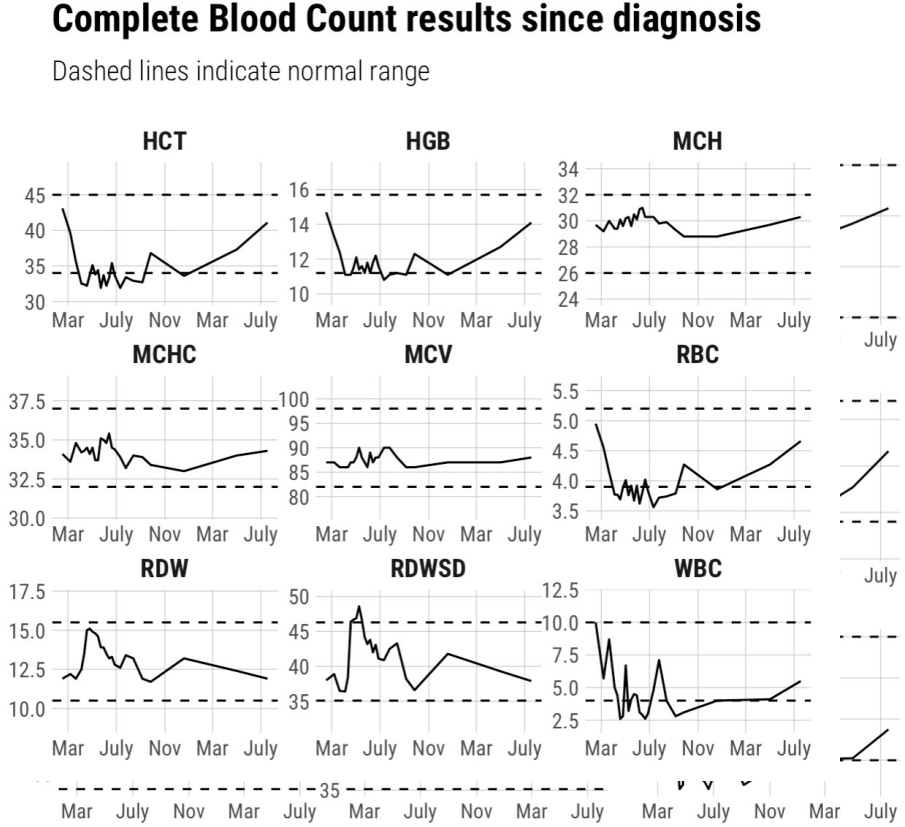
DATE	WBC 4.0-10.0 K/uL	RBC 3.9-5.2 m/uL	Hgb 11.2-15.7 g/dL	Hct 34-45 %	MCV 82-98 fL	MCH 26-32 pg	MCHC 32-37 g/dL	RDW 10.5-15.5 %	RDWSD 35.1-46.3 fL
06/13/18 8:30A (16)	2.9*	3.81*	11.8	33.3*	87	31.0	35.4	13.2	42.0
(16) TY									
06/06/18 9:00A (18)	3.1*	3.62*	11.2	32.2*	89				
(18) TY									
05/30/18 8:30A (20)	4.1	3.92	11.8	34.7*	86				
(20) TY									
05/23/18 9:00A (22)	4.5	3.67*	11.2	31.9*	87				
(22) TY									
05/16/18 8:35A (24)	4.1	3.92	11.6	34.4	88				
(24) TY									
05/09/18 9:20A (26)	3.2*	3.76*	11.4	33.8*	90				
(26) TY									
05/02/18 1:00P (28)	6.7	4.01	12.1	35.1	88				
(28) TY									
04/25/18 7:50A (30)	2.8*	3.89*	11.5	33.7*	87				
(30) TY									
04/18/18 9:20A (32)	2.6*	3.69*	11.1*	32.2*	87				
(32) TY									
04/11/18 10:00A	4.4	3.77*	11.1*	32.4*	86	29.4	34.3	13.5	38.5



It was a tedious process but I didn't have much else to do!

with a lot of `readr::parse_number()`

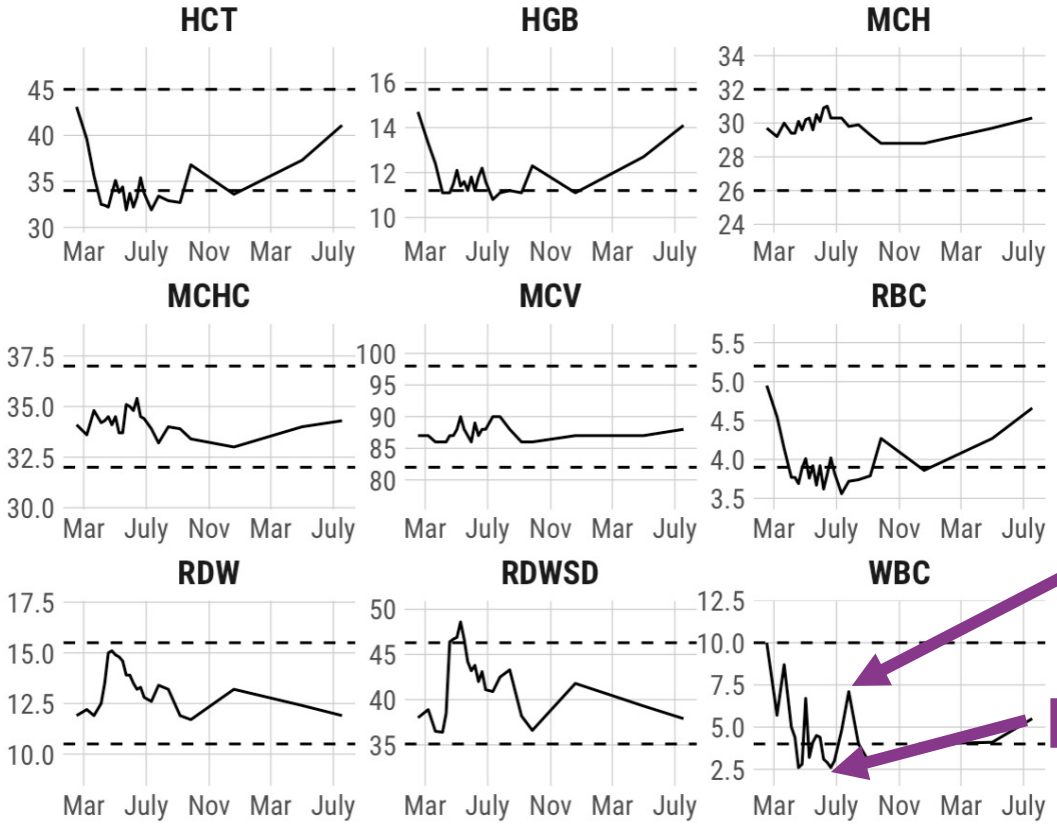
Almost…perfect



The difference? 80 lines of ggproto stuff I don't understand, from https://fishandwhistle.net/post/2018/modifying-facet-scales-in-ggplot2/

Complete Blood Count results since diagnosis

Dashed lines indicate normal range



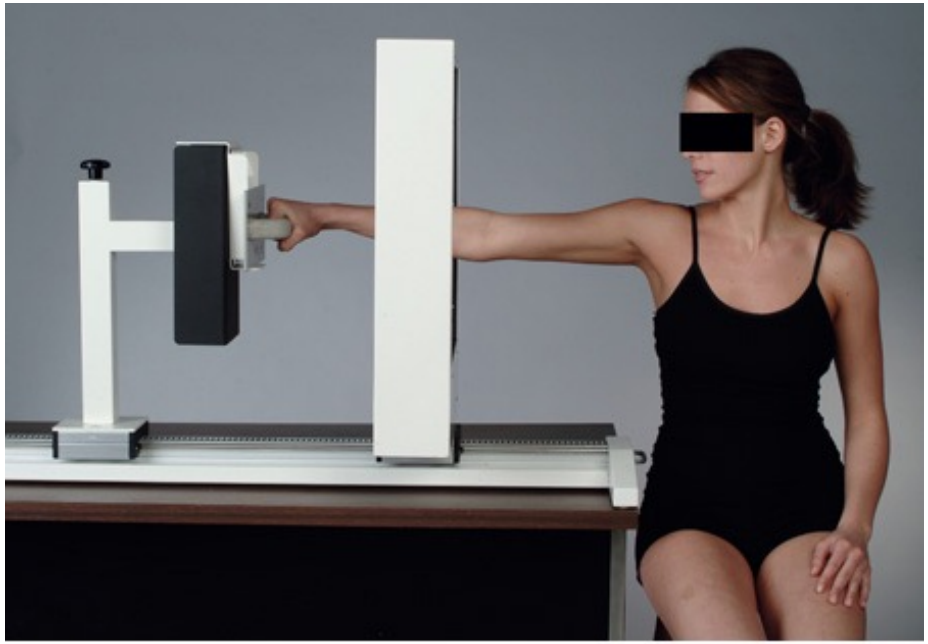
Neulasta chemo

Non-Neulasta chemo

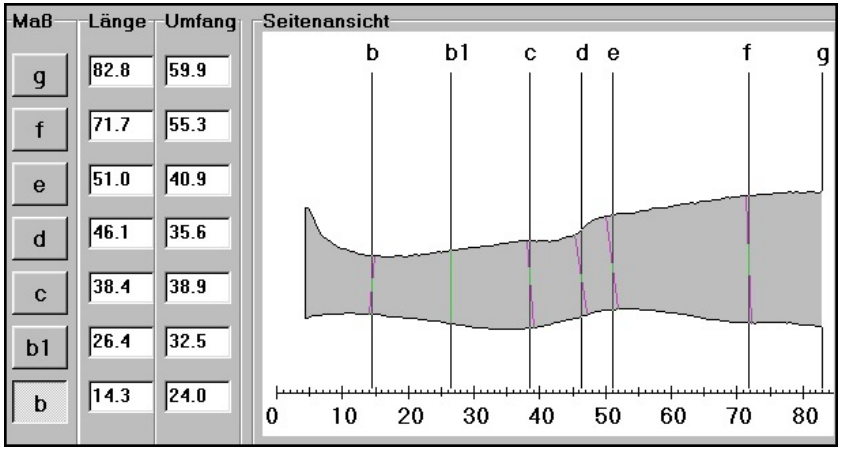
Lymphedema after axillary lymph node dissection

Image from https://columbiasurgery.org/news/2013/07/29/lympha-trial-seeks-prevent-lymphedema-breast-cancer-patients

Lymphedema monitoring



Source: Kuerer HM: *Kuerer's Breast Surgical Oncology*:
http://www.accesssurgery.com



Even better!

Image from http://www.lymphedemablog.com

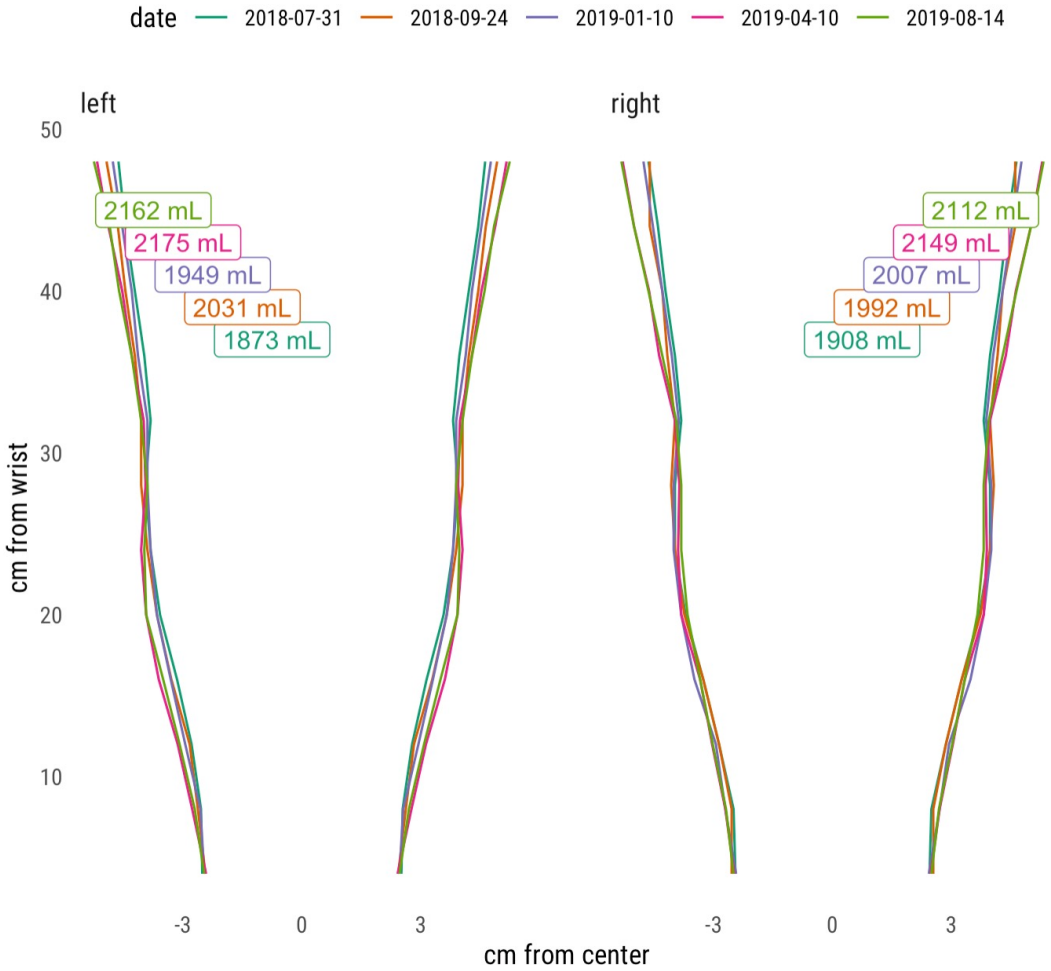
```
tribble(
         ~date,                                                  ~meas.right,
  "2018-07-31",   "15.4/15.7/18/20.5/23.3/25/25/24/25/26.5/27.7/29.2",          "15
  "2018-09-24",         "16/16/18/20.5/23.5/25/25.6/25/26.2/27/29/29",      "15.3/1
  "2019-01-10", "15.3/17/18.5/21.9/24/25.2/25.2/24.4/25.5/27/28.5/30",          "15
  "2019-04-10", "15.6/17/19.1/21/24/24.5/24.3/25/27.5/29.1/31.5/33.3", "15.2/17.4/
  "2019-08-14",     "15.7/16.9/19/21/23/24/24/24.8/27/29.2/31.5/33.5",
  )

... separate_rows() ... pivot_longer() ...

mutate(
    lag_meas = lag(meas),
    val = 4 * (meas^2 + meas * lag_meas + lag_meas^2)
  ) %>%
  summarise(new_vol = sum(val, na.rm = TRUE) / (12 * pi))
```

Formula for arm volume!

Arm measurements for lymphedema monitoring



No sign of lymphedema!

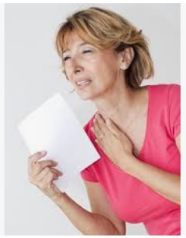
🤞🤞🤞

hot flashes and night sweats ...
health.harvard.edu

Hot flushes - causes and solutions for ...
avogel.co.uk

Hot Flashes after 60 | SheCares
shecares.com

All You Need To Know About ...
en.getmoona.com



Hot Flashes in Menop...
urmc.rochester.edu

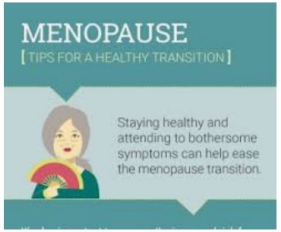
Menopausal hot flashes and night swe...
medicalnewstoday.com

Not All Hot Flashes are the Same: ...
thebiostation.com

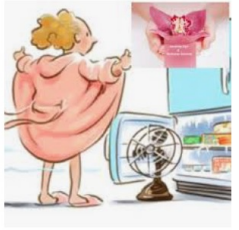
Hot Flashes - Hormonal Imbalance ...
shecares.com

Women Should Know about Hot Flashes ...
menopausenow.com

What are hot flashes and why do y...
dailywellness.com



Hot Flashes: What Can I Do?
nia.nih.gov



Hot Flashes - Gynecologist i...
serenitygyn.com

Hot Flashes, Hot Flash, Hot Flus...
renewmetoday.com

How to Tame a Hot Flash (No H...
healthywomen.org

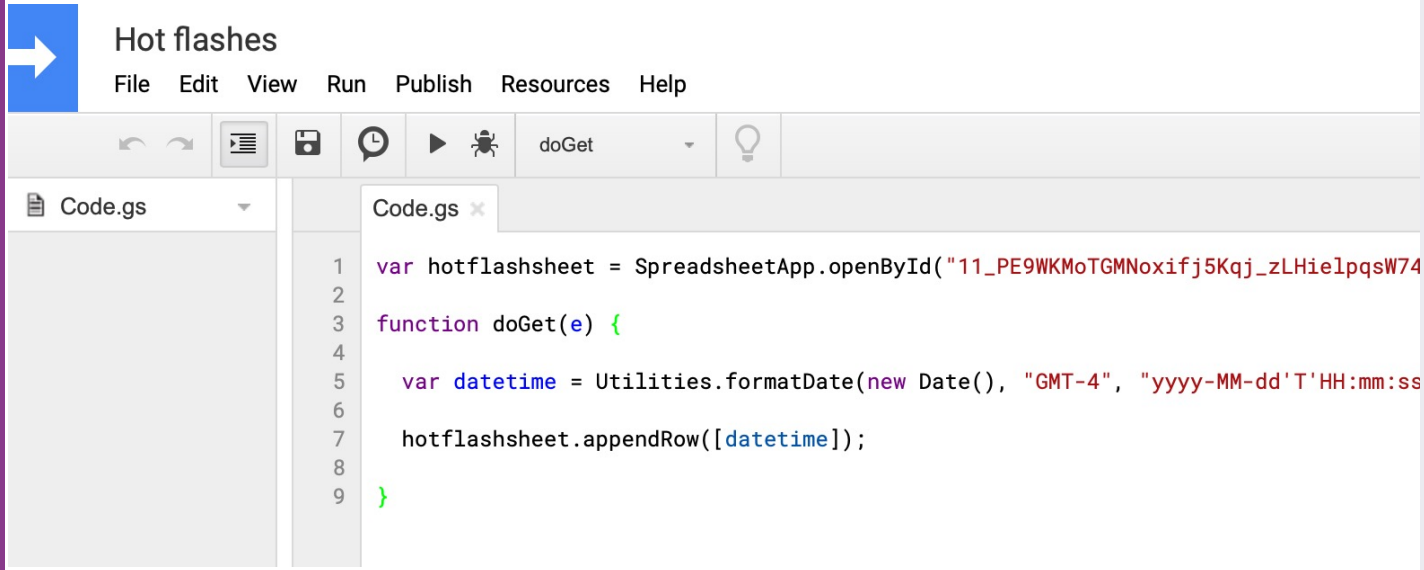
Treating hot flashes and night sweats ...
newsnetwork.mayoclinic.org

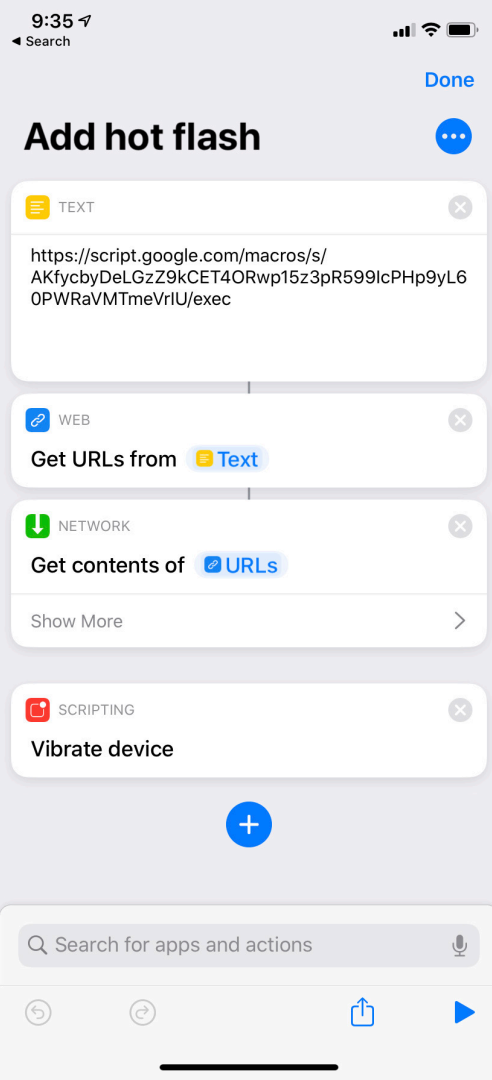
Hot Flashes Can Be Fast and Furious ...
chicagohealthonline.com

Hot Flashes Symptom Information ...
menopausenow.com

Visual Guide To Hot Flashes
webmd.com

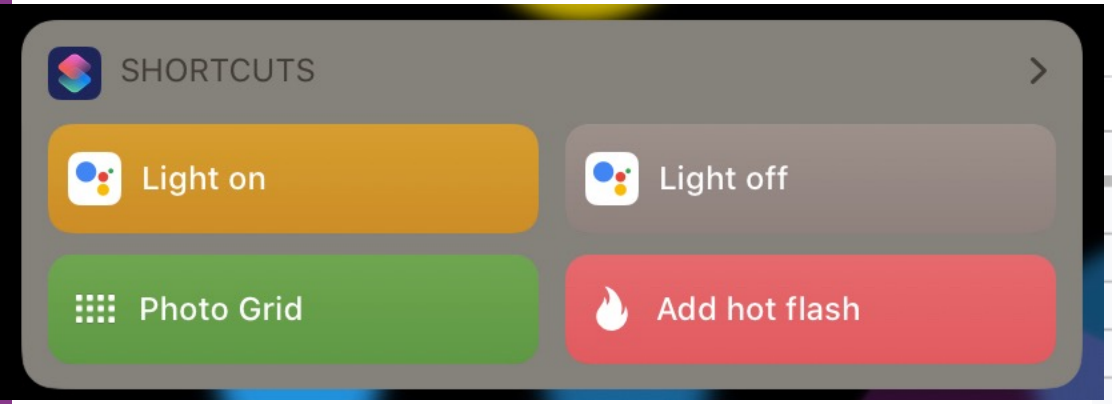
I have no idea what I'm doing….





Android option? Action Blocks?
https://www.blog.google/outreach-initiatives/accessibility/action-blocks/

But it worked!



Hot flashes

File Edit View Insert Format Data

100% $ % .0 .00

	A	B
1	datetime	
2	2019-06-27T09:42:58	
3	2019-06-27T11:04:04	
4	2019-06-27T14:29:24	
5	2019-06-27T15:16:18	
6	2019-06-27T15:51:21	
7	2019-06-27T16:33:05	
8	2019-06-27T17:23:55	
9	2019-06-27T18:09:06	
10	2019-06-27T20:57:21	

Now the R part…

Collect data

Read in data right from Google Sheets

```
hotflashes ← gs_read(gs_title("Hot flashes"),
  ws = 1, col_types = "T", verbose = FALSE
  ) %>%
  mutate(
    hour = hour(datetime),
    hour_fact = factor(hour,
      levels = 0:23,
      labels = c("midnight", paste0(1:11, "am"),
                 "noon", paste0(1:11, "pm"))
    ),
    weekday = wday(datetime, label = TRUE),
    date = date(datetime)
  )
```

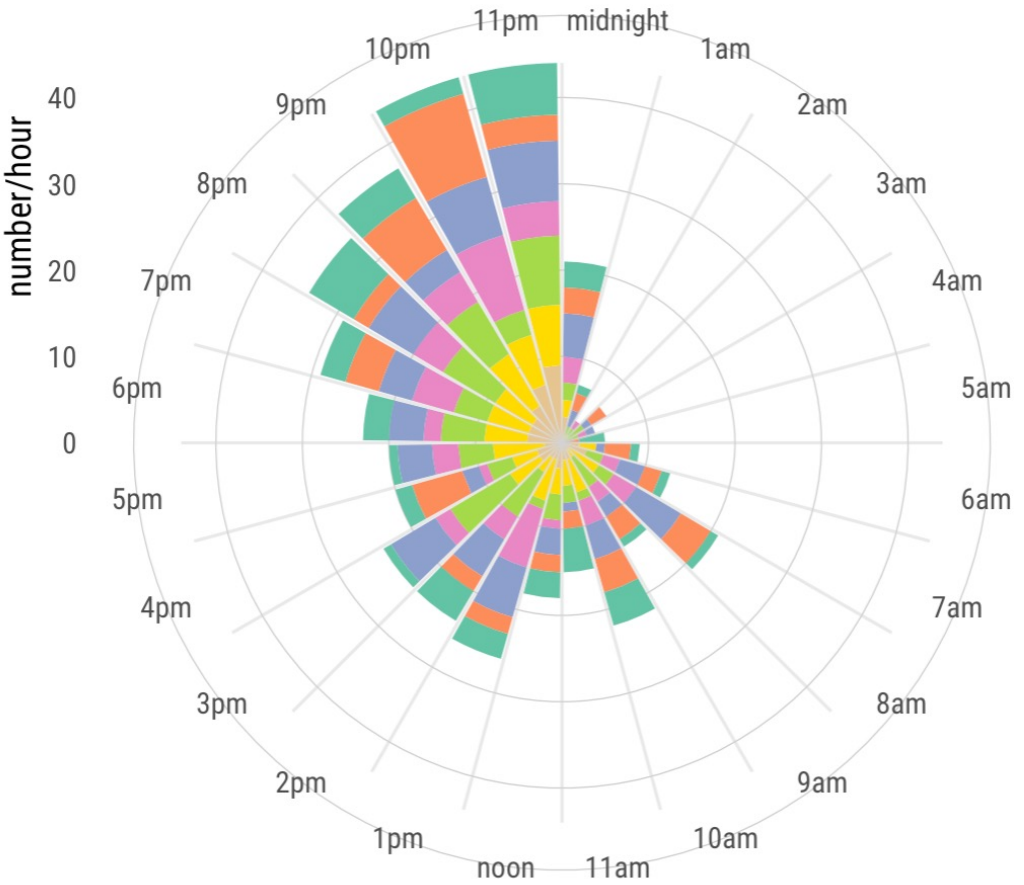
New Google API – use `googlesheets4` instead!

Visualize data

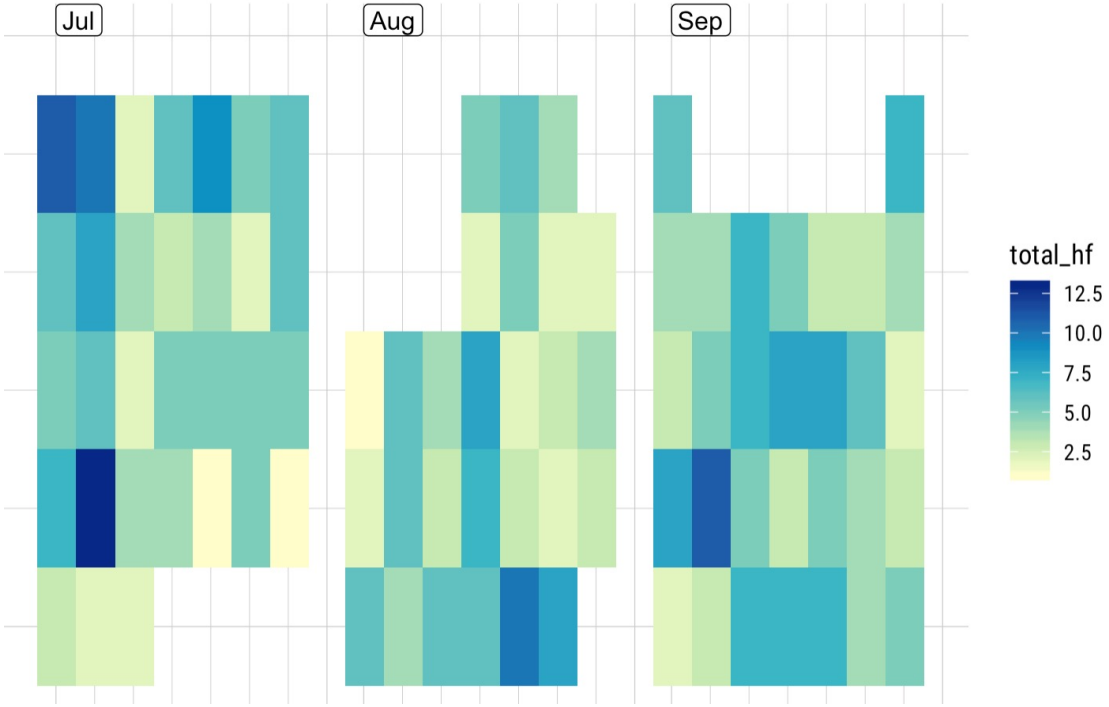
I used the `gt` package for html tables

The `sugrrants` package for time series visualization

Hot flash timing



- Worst time is bedtime
- When they wake me up I'm usually too sleepy to record
- Other missing data!



date	hotflashes
2019-06-27	10
2019-06-28	7
2019-06-29	10
2019-06-30	10
2019-07-01	11
2019-07-02	10
2019-07-03	2
2019-07-04	6
2019-07-05	9
2019-07-06	5
2019-07-07	6

I had this on a (very basic) Shiny app so I could see my data on the go!

How much was this all costing?

Me? Luckily, relatively little

https://www.npr.org/sections/health-shots/2019/02/26/696321475/cancer-complications-confusing-bills-maddening-errors-and-endless-phone-calls

My insurance company? Tonnnnnnnns

But how to get that data?

Find a Doctor & Estimate Costs

Quickly search for doctors and get cost estimates for over 1600 common medical procedures.

Review My Benefits

All of my health care info in one convenient spot.

Review My Deductible & Co-Insurance

See my current deductible, out-of-pocket max and co-insurance.

Review My Claims

Review my paid and/or pending claims.

+	08/13/2018	08/13/2018	LOUISA, SMITH	BETH ISRAEL DEACONESS MEDICALCENTER	Medical	$0.00	$15,468.38	Complete

-	08/13/2018	08/13/2018	LOUISA, SMITH	BETH ISRAEL DEACONESS MEDICALCENTER	Medical	$0.00	$15,468.38	Complete

Claim ID: 20182320587900
Date Received: 08/20/2018

Service type	What you owe	Amount your health care provider charged	Amount covered
Ancillary	**$0.00**	$1,212.00	$715.74
Ancillary	**$0.00**	$14,085.80	$4,564.12
Ancillary	**$0.00**	$67.00	$14.44
Ancillary	**$0.00**	$3.58	$0.62
Ancillary	**$0.00**	$100.00	$0.00
Total	**$0.00**	$15,468.38	$5,294.92

View the Claim Details

Process (/struggle)

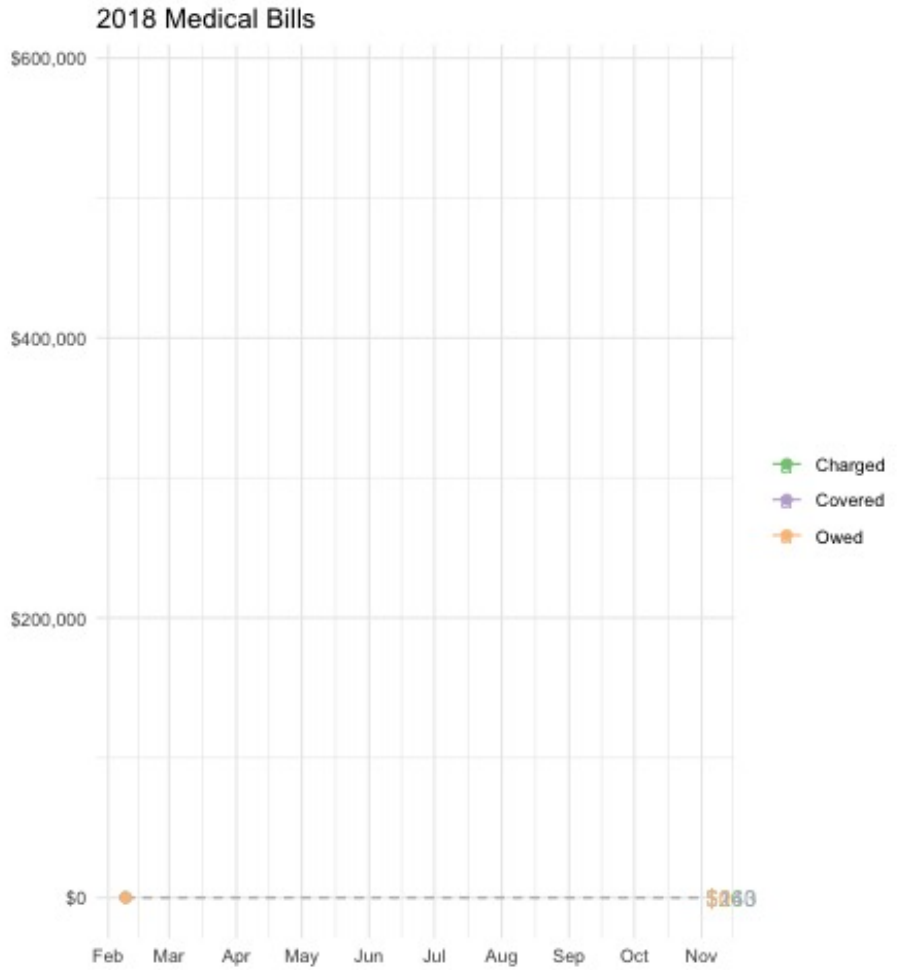
Collect data

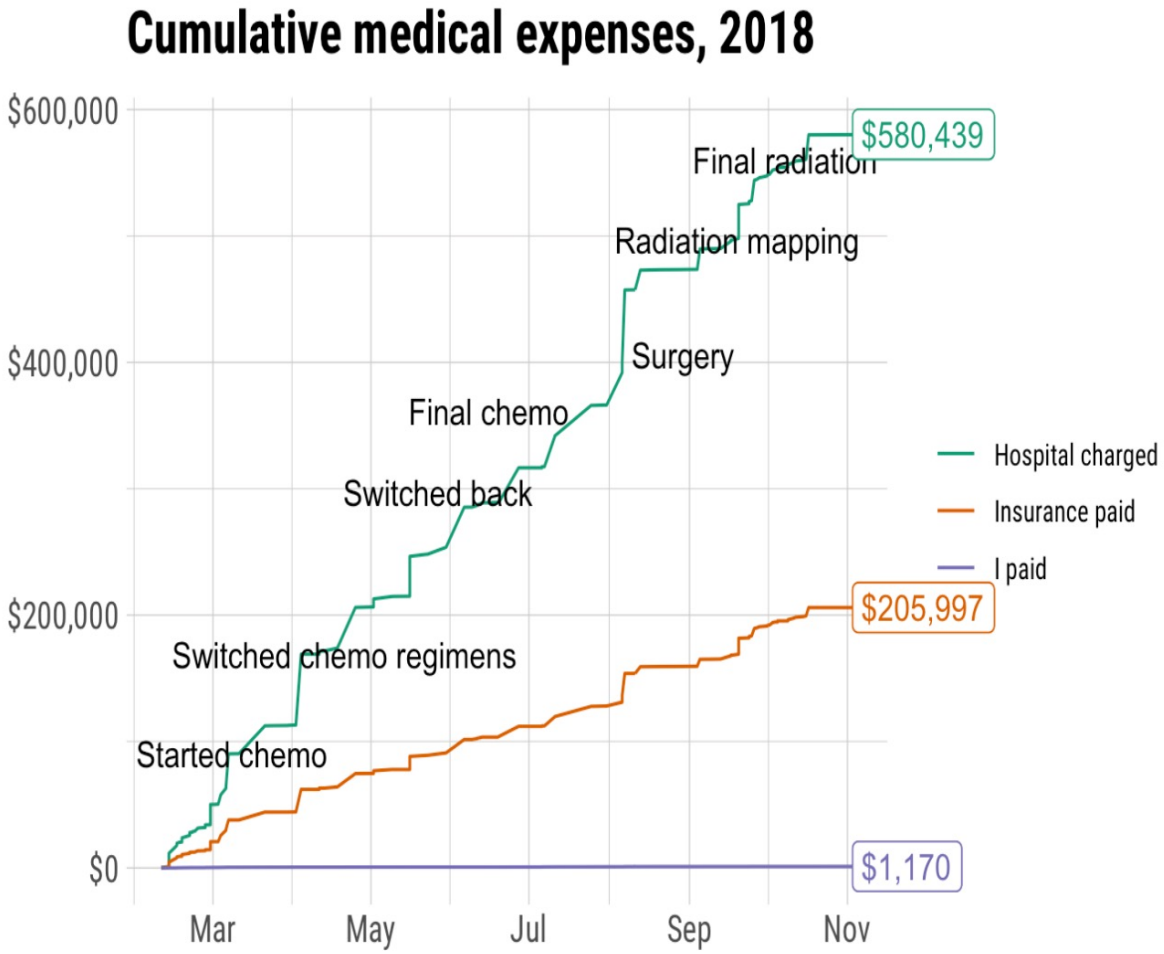
Many attempts via `rvest` to get past the password protection
Landed on `RSelenium` – allows for interactive session, easier troubleshooting (but not a lot of help out there!)
(brief demo)

Visualize data

I really wanted to make a `gganimate` gif of medical bills over time
My first ever issue filed on github!
https://github.com/thomasp85/gganimate/issues/172

Full explanation here:
https://www.louisahsmith.com/post/secrets-and-robots/


2018 Medical Bills
$600,000
$400,000
$200,000
$0
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Charged
Covered
Owed



Conclusion: (this is only medical bills, not pharmacy, but) I was really, really lucky!

Putting it all together

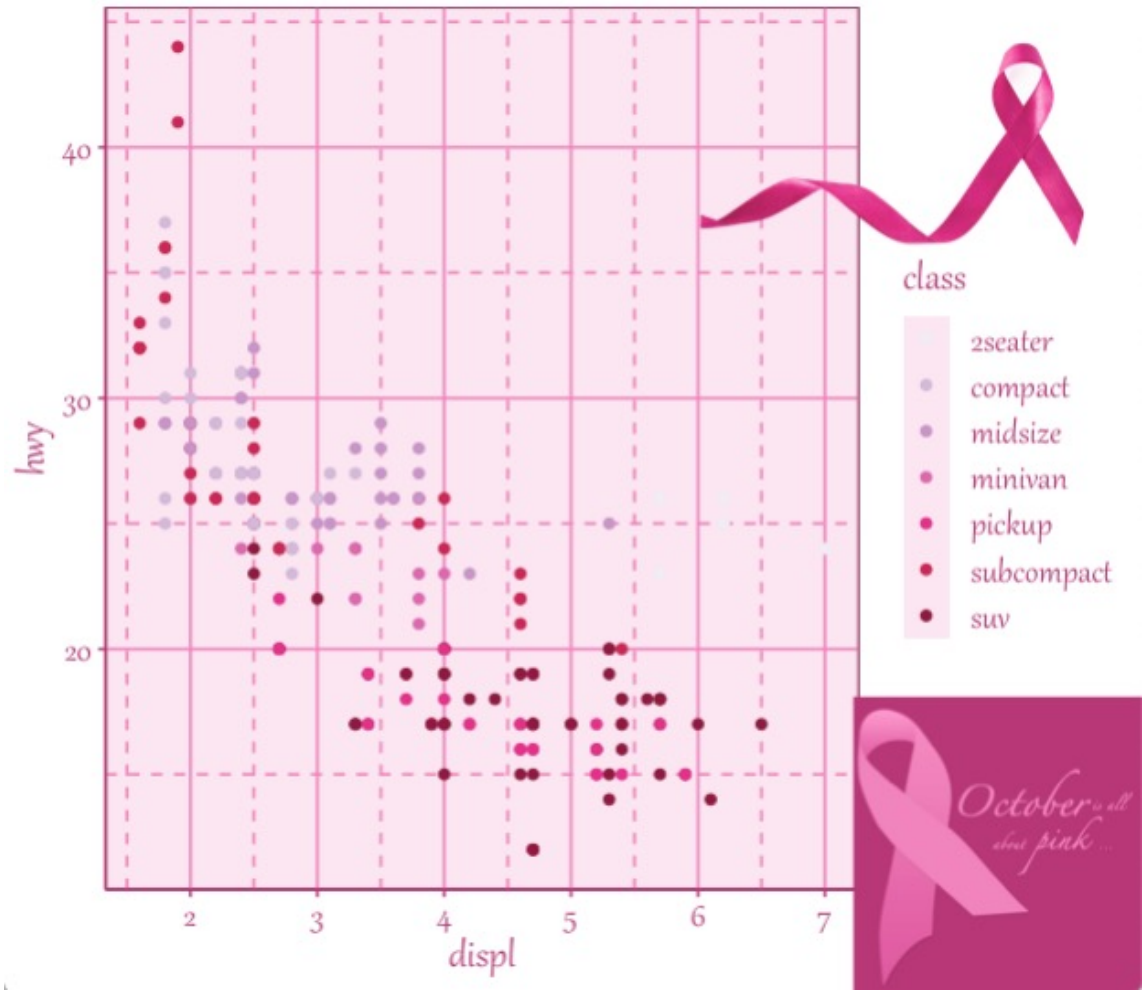
October: breast cancer awareness month

Recommended reading:

https://web.archive.org/web/20110609202708/http://www.barbaraehrenreich.com/cancerland.htm

http://thinkbeforeyoupink.org/resources/history-of-the-pink-ribbon/

https://www.nytimes.com/2015/10/31/health/breast-cancer-awareness-pink.html



If everything is going to be pink this month, why not ggplot?!

How to make a ggplot2 theme

```r
darkpink ← "#B93476"
lighterpink ← "#F282BC"
lightpink ← "#fce6f1"
purps ← RColorBrewer::brewer.pal(7, "PuRd")
```

```r
theme_bc_aware ← function() {
  darkpink ← "#B93476"
  lighterpink ← "#F282BC"
  lightpink ← "#fce6f1"
  theme_dark() %+replace%
    theme(
      title = element_text(color = darkpink, family = "Gabriola", size = rel(1.5)),
      panel.grid.major = element_line(color = lighterpink),
      panel.grid.minor = element_line(linetype = "dashed", color = lighterpink),
      panel.background = element_rect(fill = lightpink),
      panel.border = element_rect(color = darkpink, fill = NA),
      axis.line = element_line(color = darkpink),
      axis.ticks = element_line(color = darkpink),
      axis.text = element_text(color = darkpink, family = "Gabriola", size = rel(1.3)),
      strip.text = element_text(color = darkpink, family = "Gabriola", size = rel(1.3)),
      strip.background = element_rect(color = "white"),
      legend.key = element_rect(fill = lightpink, color = NA),
      legend.text = element_text(color = darkpink, family = "Gabriola", size = rel(1.3))
    )
}
```

Better resource than me: https://bookdown.org/rdpeng/RProgDA/building-a-new-theme.html

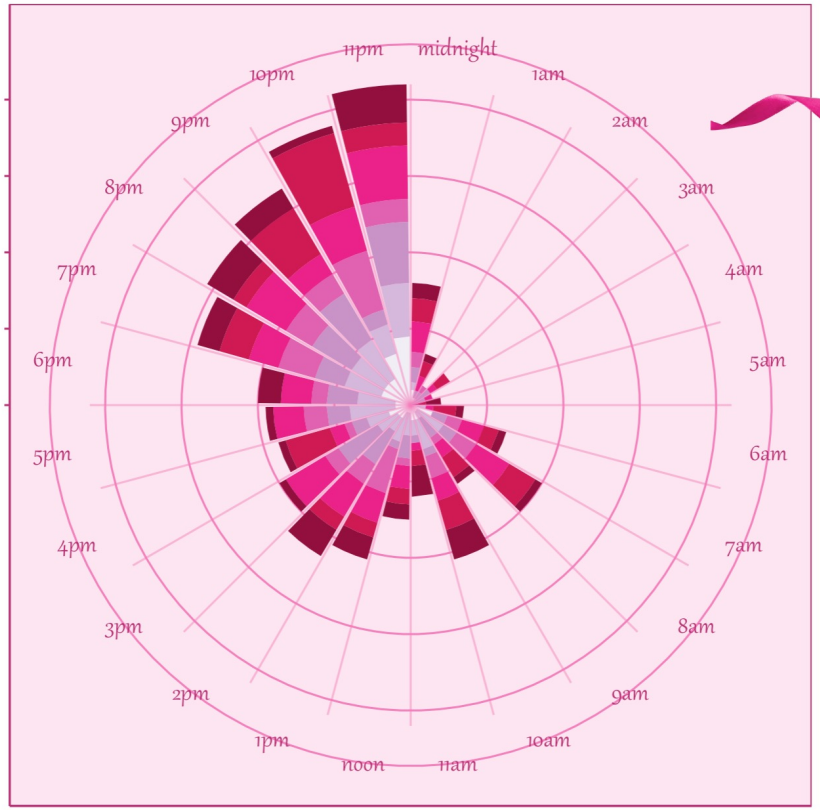
Add some logos

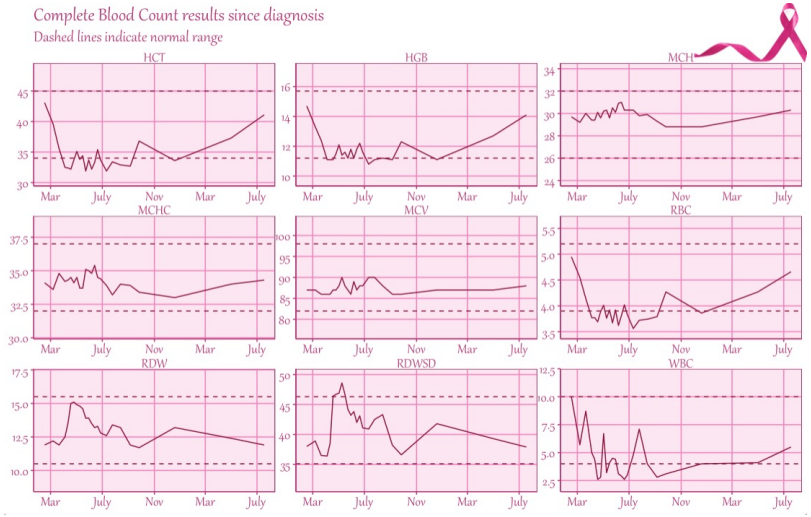
```
logo ← magick::image_read(here::here("img", "pinktober.jpg"))
ribbon ← magick::image_read(here::here("img", "ribbon.png"))
grid::grid.raster(ribbon,
  x = .95, y = .95,
  just = c("right", "top"),
  width = unit(2, "inches")
)

grid::grid.raster(logo,
  x = 1, y = 0,
  just = c("right", "bottom"),
  width = unit(1.5, "inches")
)
```

http://clipart-library.com/breast-cancer-ribbon.html

Hot flashes by time of day and day of week





https://twitter.com/louisahsmith/status/1179429494664388609

So…

My first R project was a shiny app for analyzing my running data

(way over my head but I learned A TON)

I like to collect data on myself – I know that's not true for everyone

I had a lot of time on my hands when I wasn't sick enough to lie there doing nothing but not well enough to think hard!

R packages I've mentioned using

tidyverse: https://www.tidyverse.org
lubridate: https://lubridate.tidyverse.org
datapasta: https://milesmcbain.github.io/datapasta/
RcppRoll
ggrepel: https://ggrepel.slowkow.com
googlesheets4: https://googlesheets4.tidyverse.org
gt: https://gt.rstudio.com
sugrrants: https://pkg.earo.me/sugrrants/
shiny: https://shiny.rstudio.com
Rselenium: https://ropensci.org/tutorials/rselenium_tutorial/
gganimate: https://gganimate.com

Where to find me

www.louisahsmith.com
@louisahsmith
louisa_h_smith@g.harvard.edu
Shiny app for some of my research: http://selection-bias.louisahsmith.com

I do do real work sometimes!