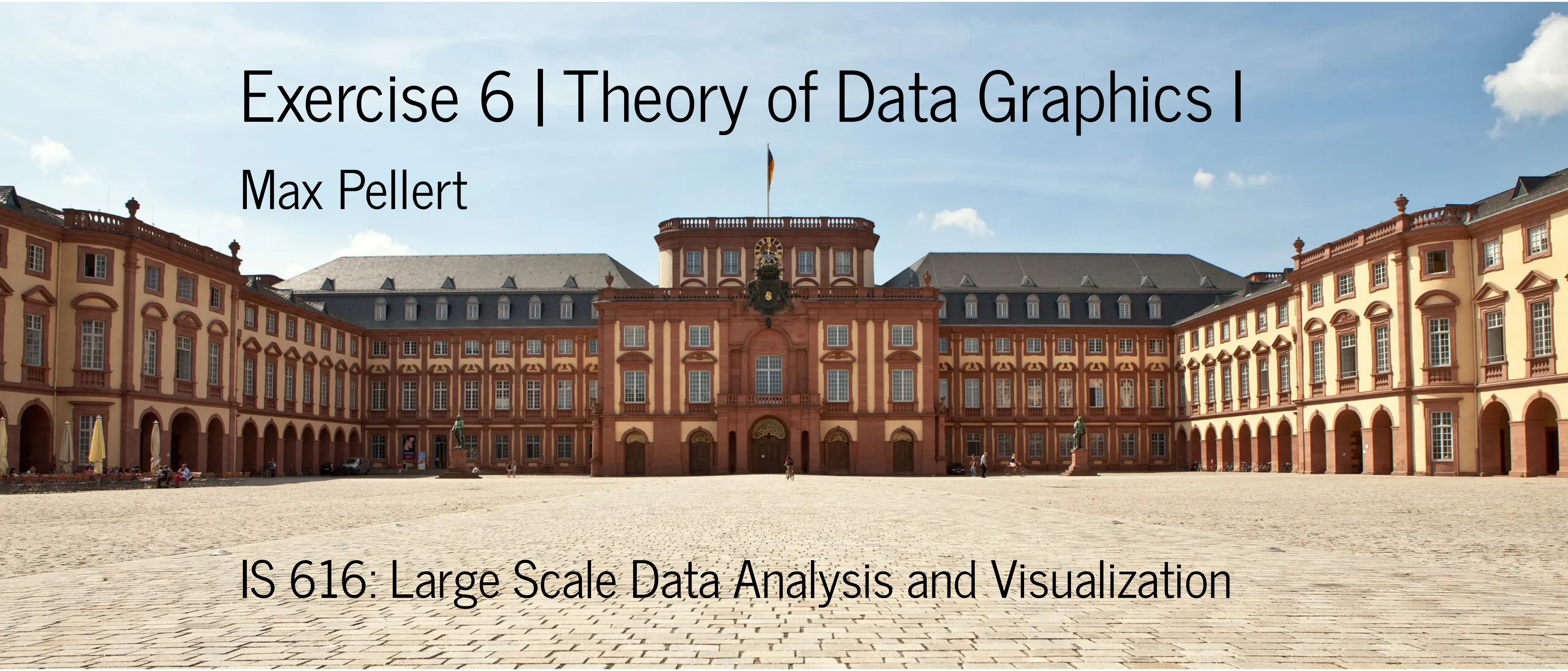
Exercise 6 | Theory of Data Graphics I

Max Pellert



IS 616: Large Scale Data Analysis and Visualization

UNIVERSITY OF MANNHEIM

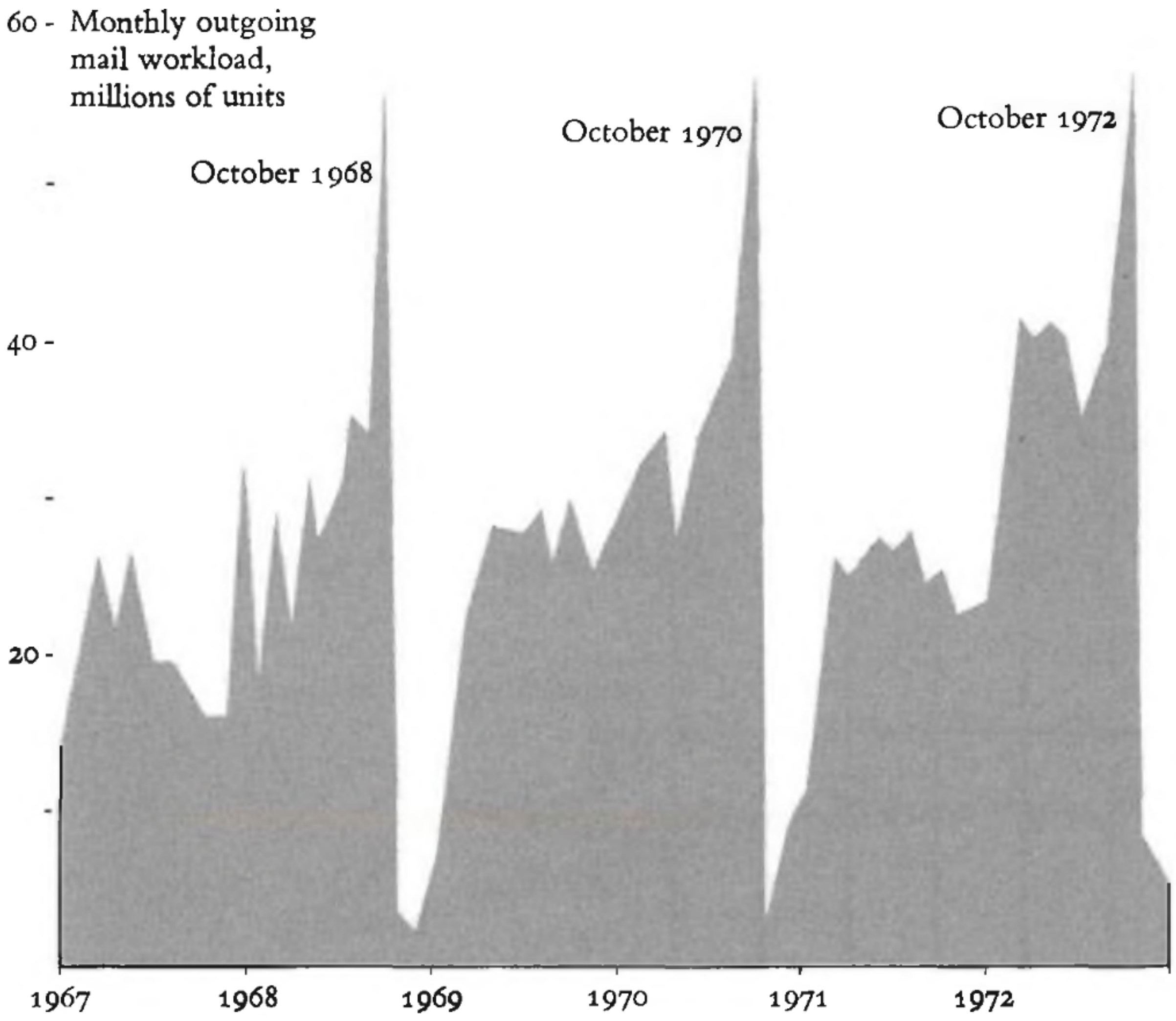
Practical considerations

We heard that we may not need graphics at all in some circumstances

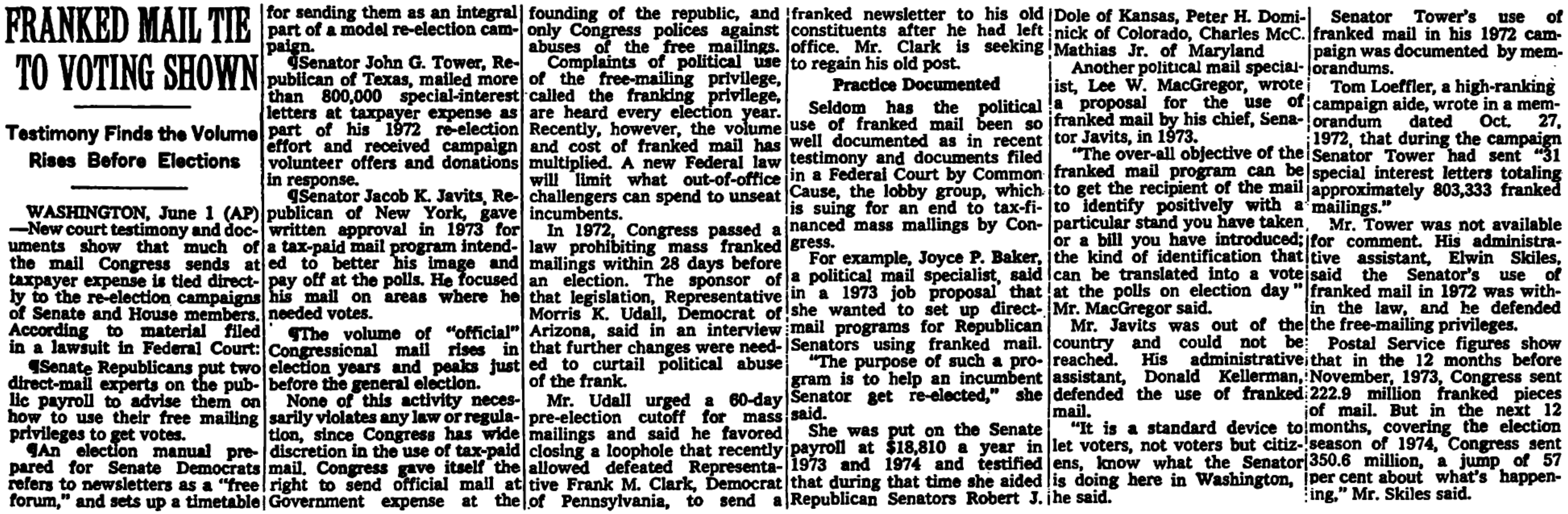
A table or just presenting the data in text can make most sense for some situations

But often we can convey much more much better with data graphics

Consider the time series about the outgoing mail of the U.S. House of Representatives that peaks every two years, just before the election day


60 -
Monthly outgoing
mail workload,
millions of units
October 1968
October 1970
October 1972
40 -
20 -
1967
1968
1969
1970
1971
1972

The graphic is worth at least 700 words, the number used in a news report describing how incumbent representatives exploit their free mailing privileges to advance their re-election campaigns:



FRANKED MAIL TIE TO VOTING SHOWN

Testimony Finds the Volume Rises Before Elections

WASHINGTON, June 1 (AP)—New court testimony and documents show that much of the mail Congress sends at taxpayer expense is tied directly to the re-election campaigns of Senate and House members. According to material filed in a lawsuit in Federal Court:

¶Senate Republicans put two direct-mail experts on the public payroll to advise them on how to use their free mailing privileges to get votes.

¶An election manual prepared for Senate Democrats refers to newsletters as a "free forum," and sets up a timetable for sending them as an integral part of a model re-election campaign.

¶Senator John G. Tower, Repubiican of Texas, mailed more than 800,000 special-interest letters at taxpayer expense as part of his 1972 re-election effort and received campaign volunteer offers and donations in response.

¶Senator Jacob K. Javits, Republican of New York, gave written approval in 1973 for a tax-paid mail program intended to better his image and pay off at the polls. He focused his mail on areas where he needed votes.

¶The volume of "official" Congressicnal mail rises in election years and peaks just before the general election.

None of this activity necessarily violates any law or regulation, since Congress has wide discretion in the use of tax-paid mail. Congress gave itself the right to send official mail at Government expense at the founding of the republic, and only Congress polices against abuses of the free mailings.

Complaints of political use of the free-mailing privilege, called the franking privilege, are heard every election year. Recently, however, the volume and cost of franked mail has multiplied. A new Federal law will limit what out-of-office challengers can spend to unseat incumbents.

In 1972, Congress passed a law prohibiting mass franked mailings within 28 days before an election. The sponsor of that legislation, Representative Morris K. Udall, Democrat of Arizona, said in an interview that further changes were needed to curtail political abuse of the frank.

Mr. Udall urged a 60-day pre-election cutoff for mass mailings and said he favored closing a loophole that recently allowed defeated Representative Frank M. Clark, Democrat of Pennsylvania, to send a franked newsletter to his old constituents after he had left office. Mr. Clark is seeking to regain his old post.

Practice Documented

Seldom has the political use of franked mail been so well documented as in recent testimony and documents filed in a Federai Court by Common Cause, the lobby group, which is suing for an end to tax-financed mass mallings by Congress.

For example, Joyce P. Baker, a political mail specialist, said in a 1973 job proposal that she wanted to set up direct-mail programs for Republican Senators using franked mail.

"The purpose of such a program is to help an incumbent Senator get re-elected," she said.

She was put on the Senate payroll at $18,810 a year in 1973 and 1974 and testified that during that time she aided Republican Senators Robert J. Dole of Kansas, Peter H. Dominick of Colorado, Charles McC. Mathias Jr. of Maryland

Another political mail specialist, Lee W. MacGregor, wrote a proposal for the use of franked mail by his chief, Senator Javits, in 1973.

"The over-all objective of the franked mail program can be to get the recipient of the mail to identify positively with a particular stand you have taken or a bill you have introduced; the kind of identification that can be translated into a vote at the polls on election day " Mr. MacGregor said.

Mr. Javits was out of the country and could not be reached. His administrative assistant, Donald Kellerman, defended the use of franked mail.

"It is a standard device to let voters, not voters but citizens, know what the Senator is doing here in Washington, he said.

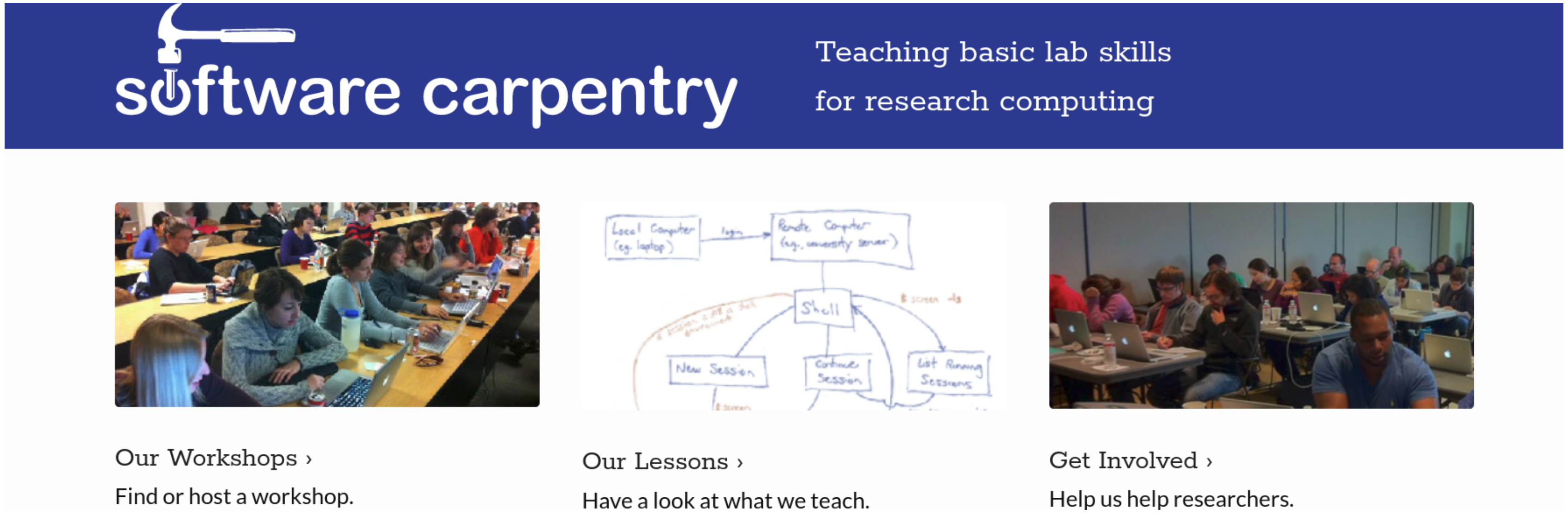
Senator Tower's use of franked mail in his 1972 campaign was documented by memorandums.

Tom Loeffler, a high-ranking campaign aide, wrote in a memorandum dated Oct. 27, 1972, that during the campaign Senator Tower had sent "31 special interest letters totaling approximately 803,333 franked mailings."

Mr. Tower was not available for comment. His administrative assistant, Elwin Skiles, said the Senator's use of franked mail in 1972 was within the law, and he defended the free-mailing privileges.

Postal Service figures show that in the 12 months before November, 1973, Congress sent 222.9 million franked pieces of mail. But in the next 12 months, covering the election season of 1974, Congress sent 350.6 million, a jump of 57 per cent about what's happening." Mr. Skiles said.

More on useful tools…



software carpentry

Teaching basic lab skills
for research computing

Local Computer (e.g. laptop)
login
Remote Computer (e.g. university server)
Shell
New Session
Continue Session
List Running Sessions

Our Workshops ›

Find or host a workshop.

Our Lessons ›

Have a look at what we teach.

Get Involved ›

Help us help researchers.

https://software-carpentry.org/

Lesson	Repository	Site
The Unix Shell	swcarpentry/shell-novice	rendered
Version Control with Git	swcarpentry/git-novice	rendered
Version Control with Mercurial	swcarpentry/hg-novice	rendered
Using Databases and SQL	swcarpentry/sql-novice-survey	rendered
Programming with Python	swcarpentry/python-novice-inflammation	rendered
Programming with R	swcarpentry/r-novice-inflammation	rendered
R for Reproducible Scientific Analysis	swcarpentry/r-novice-gapminder/	rendered
Programming with MATLAB	swcarpentry/matlab-novice-inflammation	rendered
Automation and Make	swcarpentry/make-novice	rendered
Instructor Training	carpentries/instructor-training	rendered

Software capentry

“A Software Carpentry workshop is a hands-on training that covers the core skills needed to be productive in a small research team.

Short tutorials alternate with practical exercises, and all instruction is done via live coding.”

Regularly, local workshops in many areas of the world

All lessons are also available on GitHub

https://github.com/swcarpentry/swcarpentry

More on less useful tools…

Excel recruitment time bomb makes top trainee doctors 'unappointable'

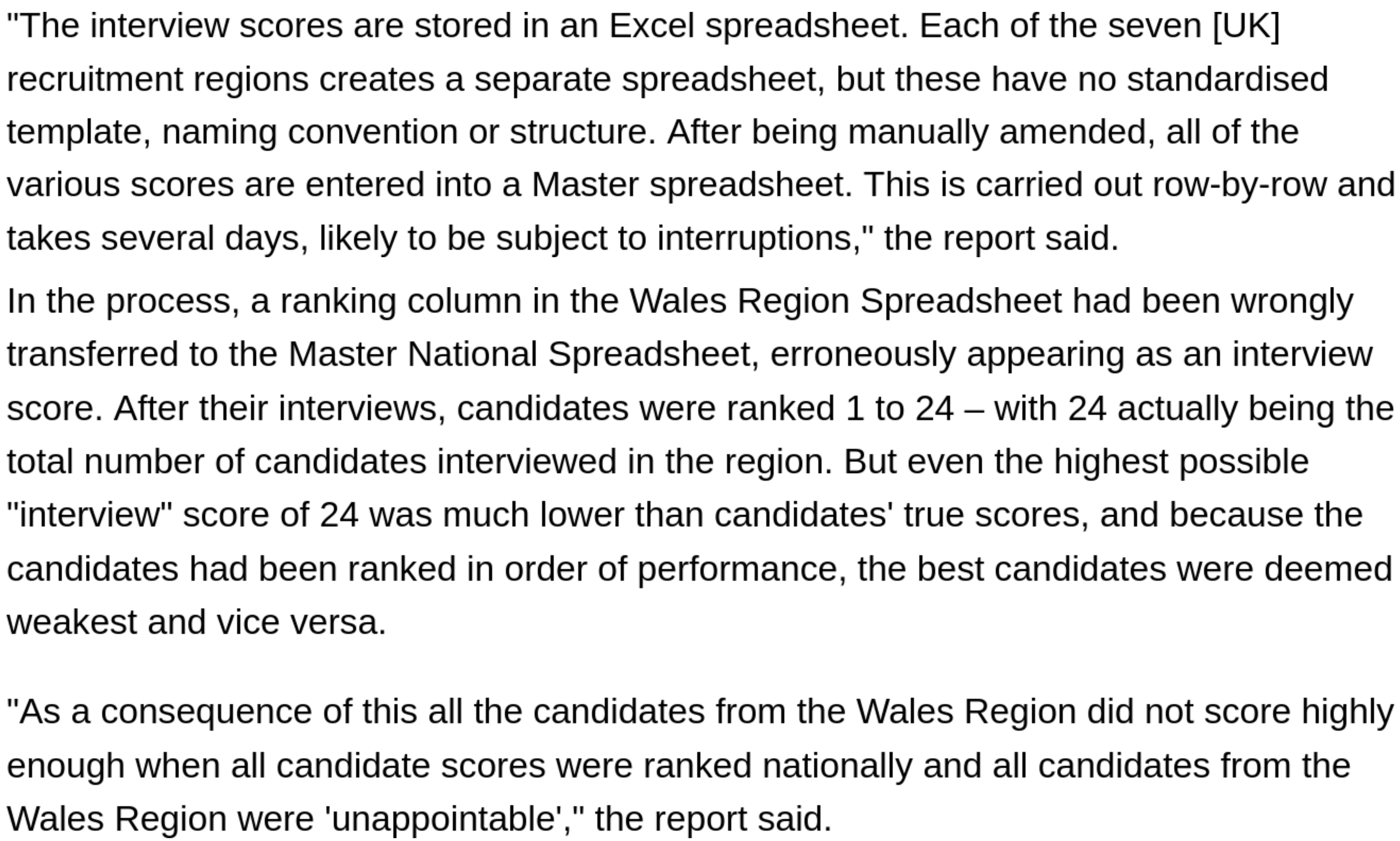
Mangled mismatch of formats, macros, and VLOOKUP practice hits wannabe anesthetists

EXCLUSIVE Computer errors, bad technology choices, and flawed processes have disrupted the recruitment of trainee anesthetists in England and Wales.

In autumn 2021, candidates seeking their third-level specialist training position (ST3) were looking forward to hearing where they would end up in one of the NHS's most sought-after medical disciplines.

However, the body responsible for their selection and recruitment – the Anaesthetic National Recruitment Office (ANRO) – told all the candidates for positions in Wales they were "unappointable," despite some of them achieving the highest interview scores.

Only when one of the candidates challenged the decision did ANRO realize its error. A subsequent Significant Incident Review showed a complex and confused approach to using spreadsheets led to the disaster.



"The interview scores are stored in an Excel spreadsheet. Each of the seven [UK] recruitment regions creates a separate spreadsheet, but these have no standardised template, naming convention or structure. After being manually amended, all of the various scores are entered into a Master spreadsheet. This is carried out row-by-row and takes several days, likely to be subject to interruptions," the report said.

In the process, a ranking column in the Wales Region Spreadsheet had been wrongly transferred to the Master National Spreadsheet, erroneously appearing as an interview score. After their interviews, candidates were ranked 1 to 24 – with 24 actually being the total number of candidates interviewed in the region. But even the highest possible "interview" score of 24 was much lower than candidates' true scores, and because the candidates had been ranked in order of performance, the best candidates were deemed weakest and vice versa.

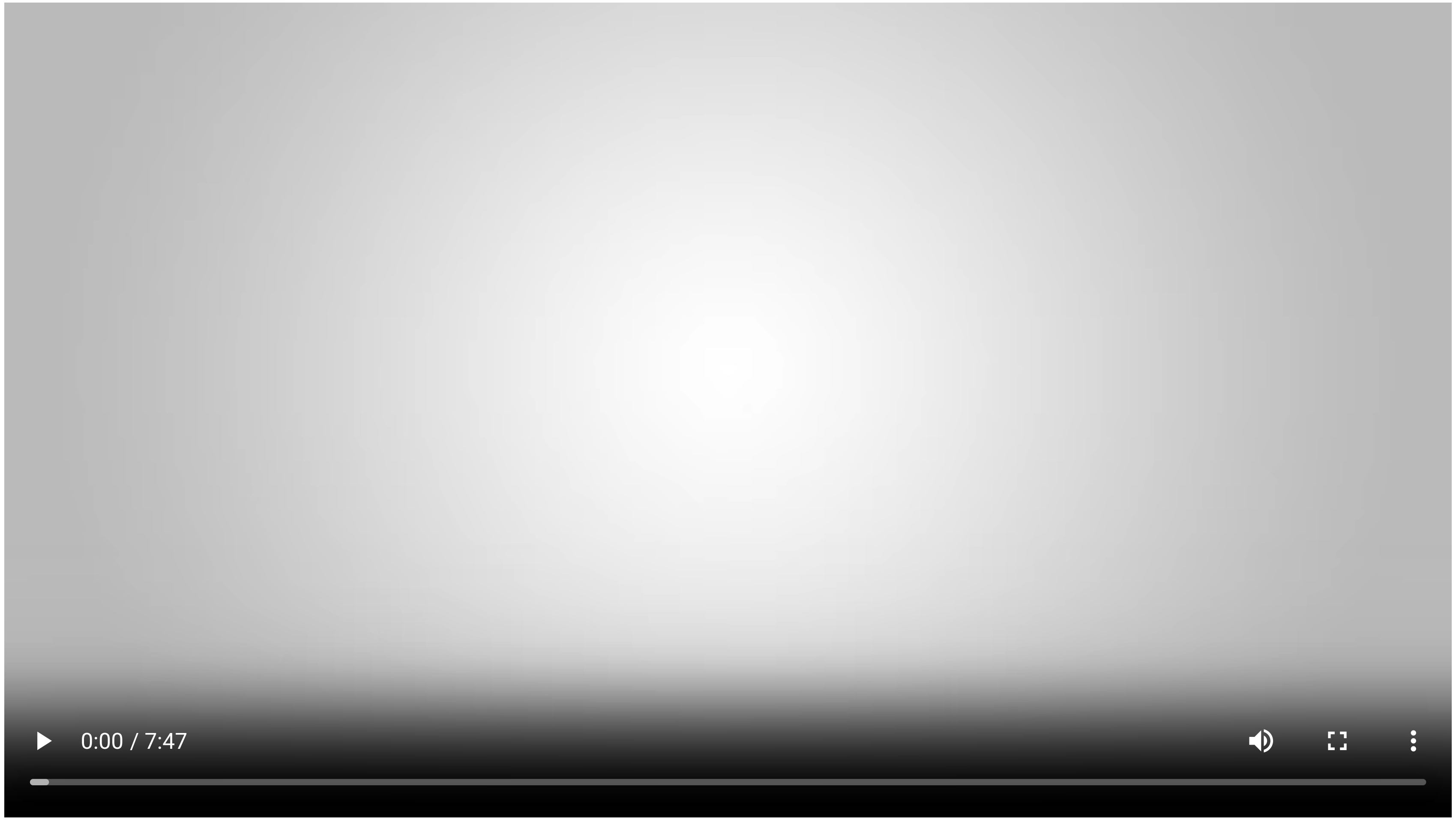
"As a consequence of this all the candidates from the Wales Region did not score highly enough when all candidate scores were ranked nationally and all candidates from the Wales Region were 'unappointable'," the report said.

https://www.theregister.com/2023/10/12/excel_anesthetist_recruitment_blunder/

In attempting to tell candidates about the problems with the scoring system, ANRO then found a bug in the messaging system of Oriel, the recruitment platform from vendor HiCom.

Oh sorry, we meant...

ANRO decided to honor the 10 job offers it had made by mistake and used Oriel to tell the candidates. Unfortunately, a system error in Oriel meant it then erroneously sent that communication to an *additional* 16 candidates. ANRO decided to honor these 16 additional offers too, and find the candidates posts.


0:00 / 7:47

Hand-in Exercise

Hand-in Exercise

The first of two that have to be completed to be able to take part in the exam

Due until October, 23rd 23:59 (AoE)

To be handed in on ILIAS (upload form provided there)

Everybody works on it on their own and uploads it individually (again, necessary for exam!)

I do check for substantial overlaps between your handed-in materials (code, visualization and text) and those of other students

Hand-in Exercise Format

Your submission should, in that order, consist of three parts

1. Your visualization (as a vector graphic!), that can also be made up of multiple sub-plots together with annotations for example
2. The documented code that produces your visualization (each line commented), R or Python
3. Half a page (A4) of explanation and reasoning for design choices that you took, what questions you wanted to answer and also explain how you structured the data for your chosen visualization and if you faced challenges and how you overcame them in case

Hand-in Exercise Format

Submit your solution as 1 (!) PDF with a filename that includes your name and additionally includes a personal identifier of you (name and student number) on every A4 page in the PDF document

If you have seperate PDFs, you can for example combine them with the following command line tool

```
pdftk file1.pdf file2.pdf file3.pdf cat output first_submission_name.pdf
```

Or use any other tool of your choice (also consider creating your document directly in Rmarkdown or IPython notebooks)

Hand-in Exercise Data

This data comes from Hollywood Age Gap via Data Is Plural:

An informational site showing the age gap between movie love interests.

The data follows certain rules:

- The two (or more) actors play actual love interests (not just friends, coworkers, or some other non-romantic type of relationship)
- The youngest of the two actors is at least 17 years old
- Not animated characters

Hand-in Exercise Data

“Note: The age gaps dataset includes”gender” columns, which always contain the values “man” or “woman”. These values appear to indicate how the *characters* in each film identify. Some of these values do not match how the *actor* identifies. We apologize if any characters are misgendered in the data!!”

https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-02-14/readme.md#hollywood-age-gaps

https://hollywoodagegap.com/

https://www.data-is-plural.com/archive/2018-02-07-edition/

age_gaps.csv

variable	class	description
movie_name	character	Name of the film
release_year	integer	Release year
director	character	Director of the film
age_difference	integer	Age difference between the characters in whole years
couple_number	integer	An identifier for the couple in case multiple couples are listed for this film
actor_1_name	character	The name of the older actor in this couple
actor_2_name	character	The name of the younger actor in this couple

age_gaps.csv

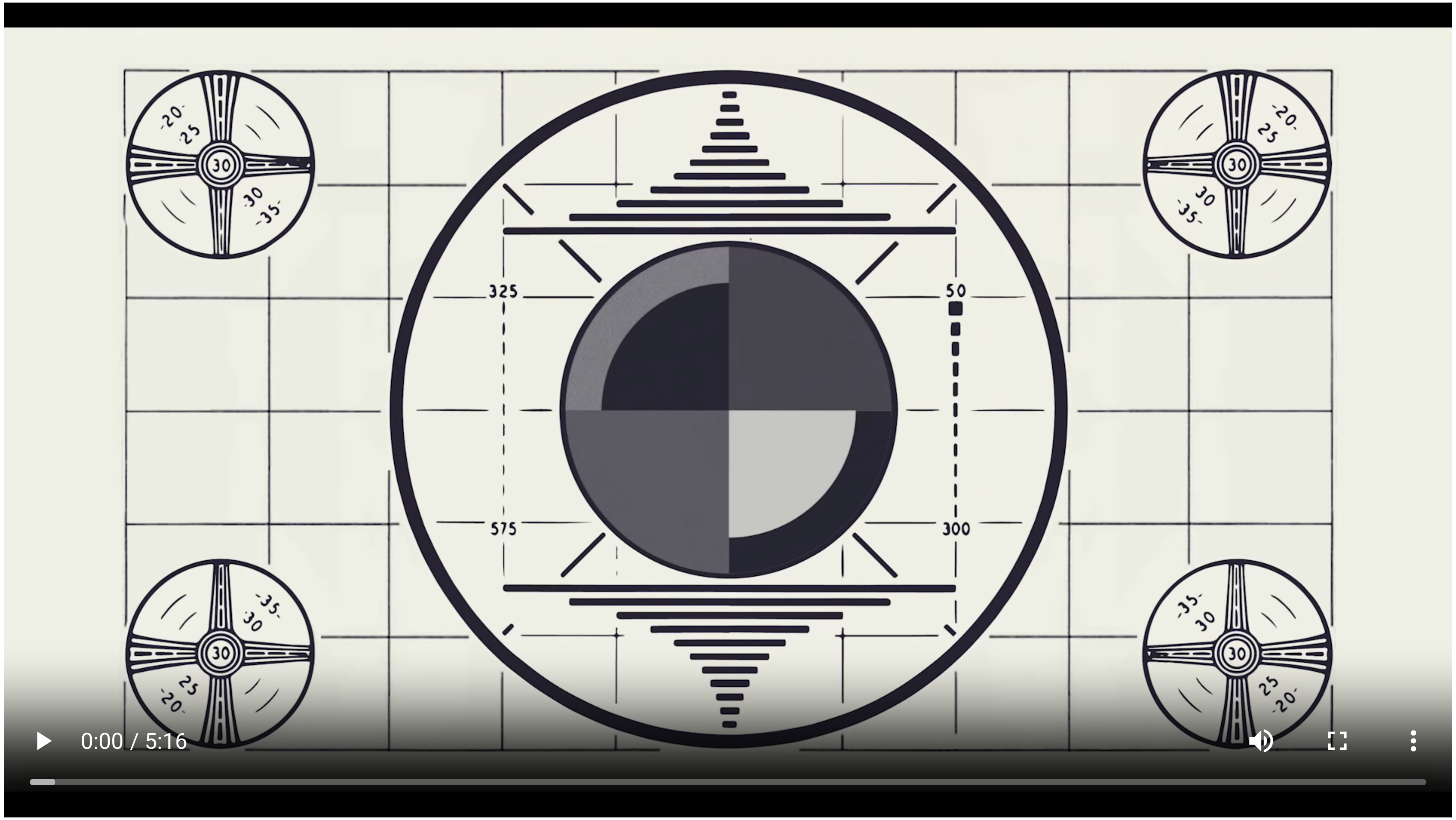
variable	class	description
character_1_gender	character	The gender of the older character, as identified by the person who submitted the data for this couple
character_2_gender	character	The gender of the younger character, as identified by the person who submitted the data for this couple
actor_1_birthdate	date	The birthdate of the older member of the couple
actor_2_birthdate	date	The birthdate of the younger member of the couple
actor_1_age	integer	The age of the older actor when the film was released
actor_2_age	integer	The age of the younger actor when the film was released

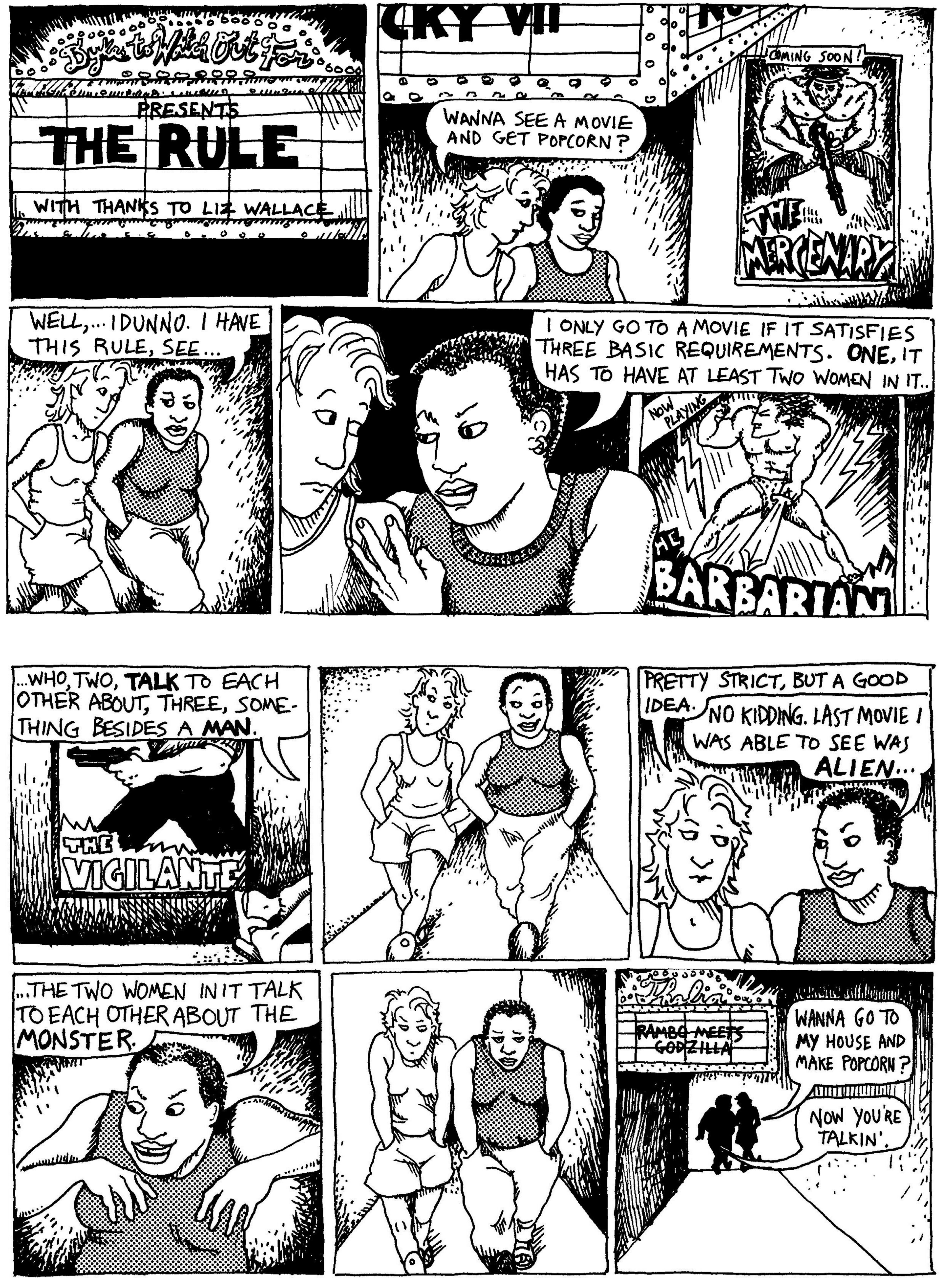
`age_gaps.csv`

https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-02-14/age_gaps.csv

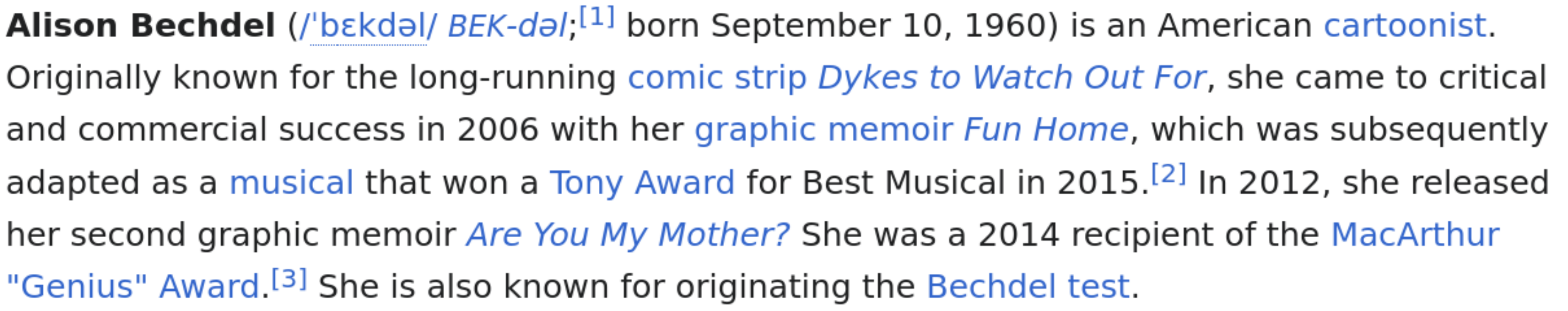
`age_gaps.csv` in R

```r
# Get the Data

# Read in with tidytuesdayR package
# Install from CRAN via: install.packages("tidytuesdayR")
# This loads the readme and all the datasets for the week of interest

# Either ISO-8601 date or year/week works!

tuesdata <- tidytuesdayR::tt_load('2023-02-14')
tuesdata <- tidytuesdayR::tt_load(2023, week = 7)

age_gaps <- tuesdata$age_gaps

# Or read in the data manually

age_gaps <- readr::read_csv('https://raw.githubusercontent.com/rfordatas
```

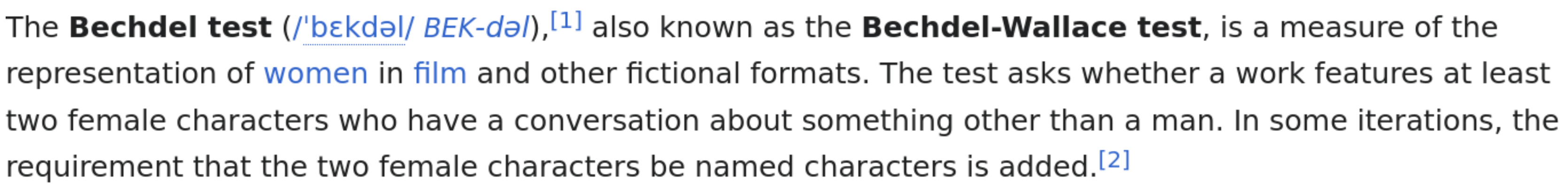

20
25
30
35
325
50
575
300
0:00 / 5:16


Dykes to Watch Out For
PRESENTS
THE RULE
WITH THANKS TO LIZ WALLACE
CRY VII
WANNA SEE A MOVIE AND GET POPCORN?
COMING SOON!
THE MERCENARY
WELL,... I DUNNO. I HAVE THIS RULE, SEE...
I ONLY GO TO A MOVIE IF IT SATISFIES THREE BASIC REQUIREMENTS. **ONE**, IT HAS TO HAVE AT LEAST TWO WOMEN IN IT..
NOW PLAYING
THE BARBARIAN
...WHO, TWO, **TALK** TO EACH OTHER ABOUT, THREE, SOMETHING BESIDES A **MAN**.
THE VIGILANTE
PRETTY STRICT, BUT A GOOD IDEA.
NO KIDDING. LAST MOVIE I WAS ABLE TO SEE WAS **ALIEN**...
...THE TWO WOMEN IN IT TALK TO EACH OTHER ABOUT THE **MONSTER**.
Thalia
RAMBO MEETS GODZILLA
WANNA GO TO **MY** HOUSE AND MAKE POPCORN?
NOW YOU'RE TALKIN'.

Bechdel Test



Alison Bechdel (/ˈbɛkdəl/ *BEK-dəl*;[1] born September 10, 1960) is an American cartoonist. Originally known for the long-running comic strip *Dykes to Watch Out For*, she came to critical and commercial success in 2006 with her graphic memoir *Fun Home*, which was subsequently adapted as a musical that won a Tony Award for Best Musical in 2015.[2] In 2012, she released her second graphic memoir *Are You My Mother?* She was a 2014 recipient of the MacArthur "Genius" Award.[3] She is also known for originating the Bechdel test.



The **Bechdel test** (/ˈbɛkdəl/ *BEK-dəl*),[1] also known as the **Bechdel-Wallace test**, is a measure of the representation of women in film and other fictional formats. The test asks whether a work features at least two female characters who have a conversation about something other than a man. In some iterations, the requirement that the two female characters be named characters is added.[2]

The Dollar-And-Cents Case Against Hollywood's Exclusion of Women

https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

Bechdel Test

“We previously provided a dataset about the Bechdel Test. It might be interesting to see whether there is any correlation between these datasets! The Bechdel Test dataset also included additional information about the films that were used in that dataset.”

https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-03-09/readme.md

raw_bechdel.csv

variable	class	description
year	integer	Year of release
id	integer	ID of film
imdb_id	character	IMDB ID
title	character	Title of film
rating	integer	Rating (0-3), 0 = unscored, 1. It has to have at least two [named] women in it, 2. Who talk to each other, 3. About something besides a man

movies.csv

variable	class	description
year	double	Year
imdb	character	IMDB
title	character	Title of movie
test	character	Bechdel Test outcome
clean_test	character	Bechdel Test cleaned
binary	character	Binary pass/fail of bechdel
budget	double	Budget as of release year
domgross	character	Domestic gross in release year
intgross	character	International gross in release year
code	character	Code

movies.csv

variable	class	description
budget_2013	double	Budget normalized to 2013
domgross_2013	character	Domestic gross normalized to 2013
intgross_2013	character	International gross normalized to 2013
period_code	double	Period code
decade_code	double	Decade Code
imdb_id	character	IMDB ID
plot	character	Plot of movie
rated	character	Rating of movie
response	character	Response?
language	character	Language of film
country	character	Country produced in
writer	character	Writer of film

movies.csv

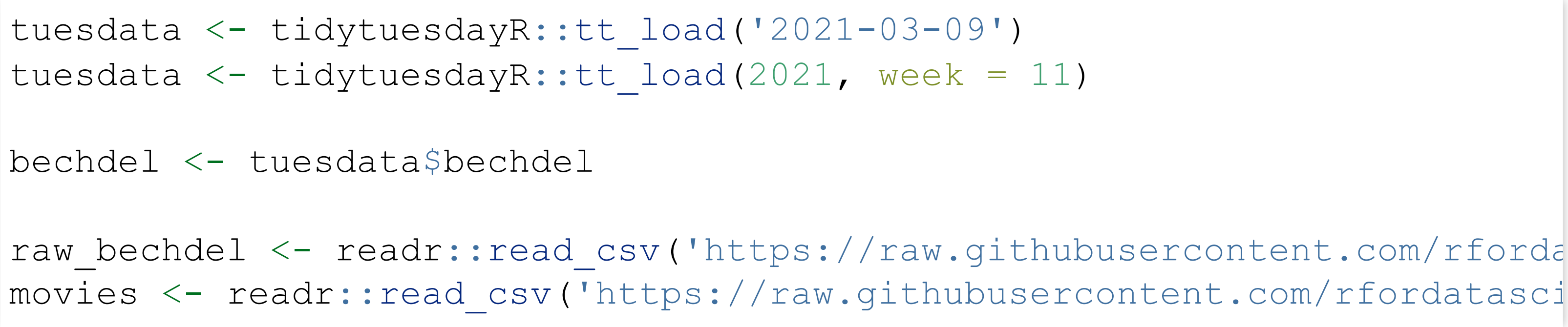
variable	class	description
metascore	double	Metascore rating (0-100)
imdb_rating	double	IMDB Rating 0-10
director	character	Director of movie
released	character	Released date
actors	character	Actors
genre	character	Genre
awards	character	Awards
runtime	character	Runtime
type	character	Type of film
poster	character	Poster image
imdb_votes	character	IMDB Votes
error	character	Error?

raw_bechdel.csv & movies.csv

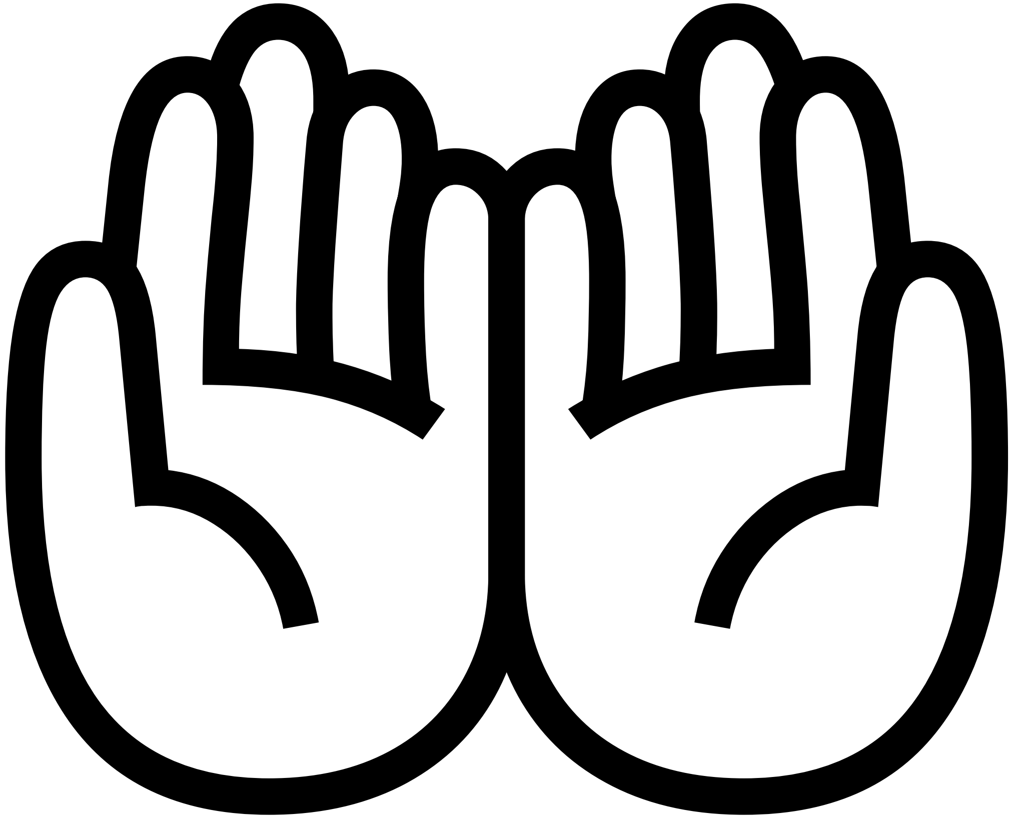
https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-03-09/raw_bechdel.csv

https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-03-09/movies.csv

raw_bechdel.csv & movies.csv in R



```
tuesdata <- tidytuesdayR::tt_load('2021-03-09')
tuesdata <- tidytuesdayR::tt_load(2021, week = 11)

bechdel <- tuesdata$bechdel

raw_bechdel <- readr::read_csv('https://raw.githubusercontent.com/rforda
movies <- readr::read_csv('https://raw.githubusercontent.com/rfordatasci
```



Acknowledgements

https://www.youtube.com/watch?v=AdSZJzb-aX8#!

https://www.youtube.com/watch?v=Meq3CyuKOjM

https://yy.github.io/dviz-course/