A gentle introduction to R

Neil Ketchley

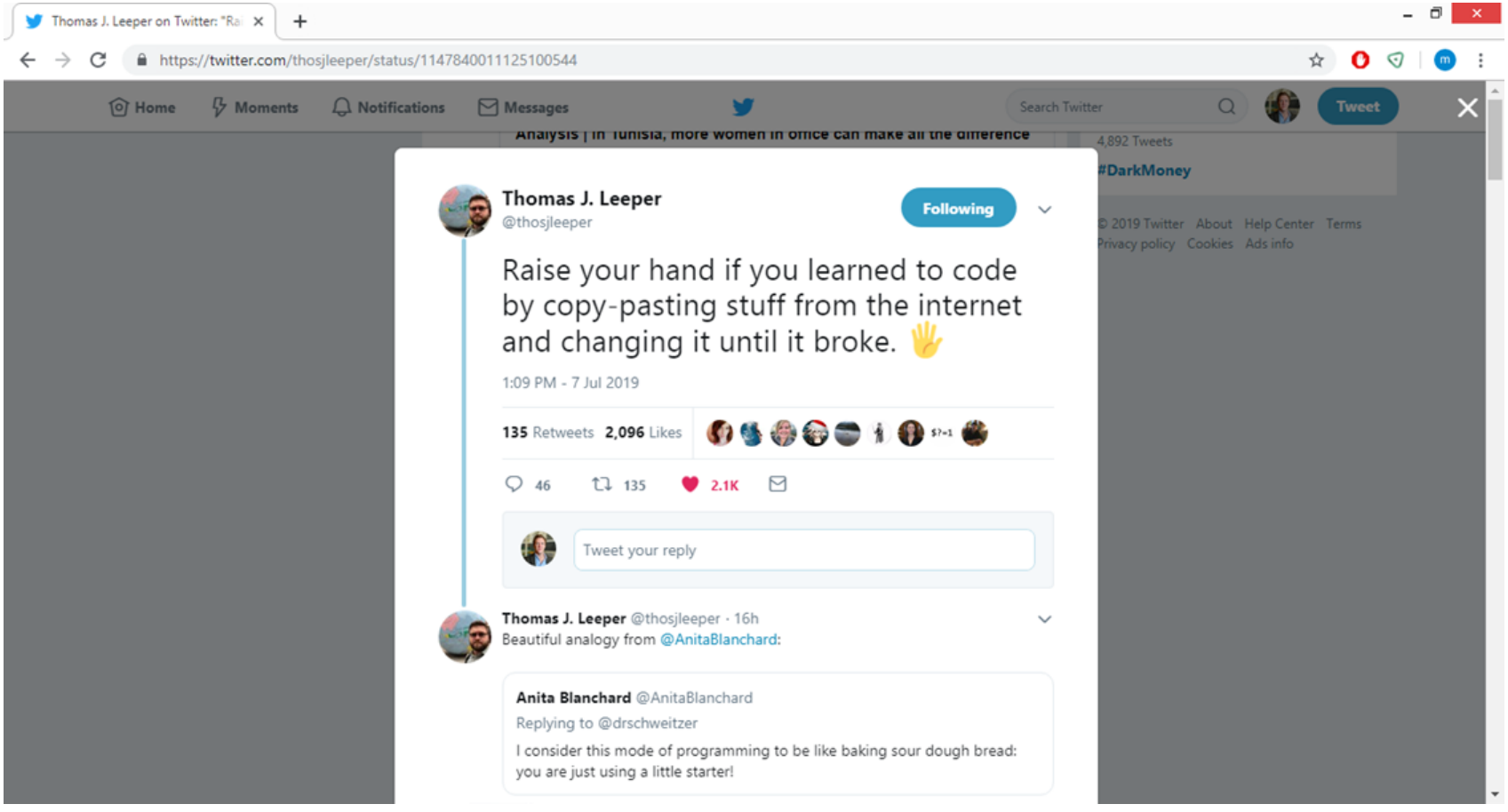
Professor of Political Science, UiO

Senior Research Fellow, NUPI

ketchley@nupi.no

05/26/21

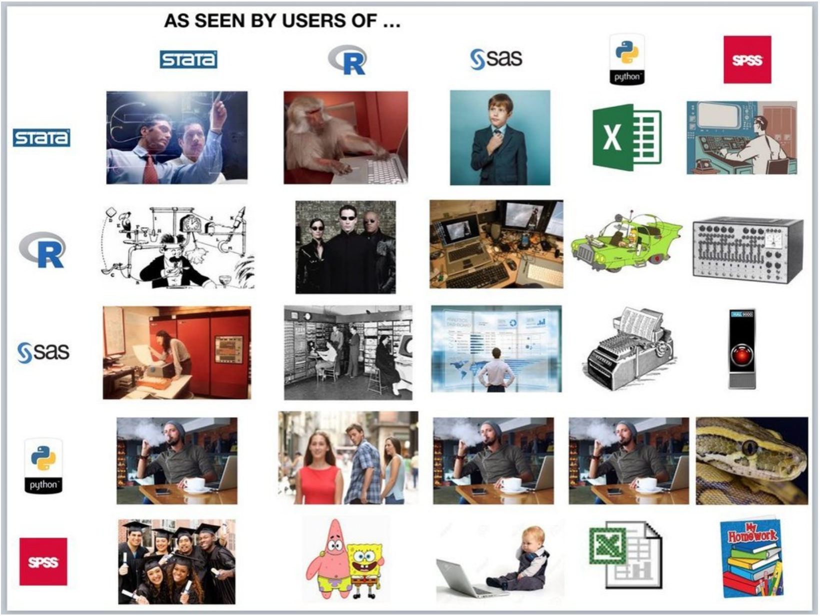
Most people aren't natural programmers



Check out Thomas Leeper: https://thomasleeper.com/

Topics

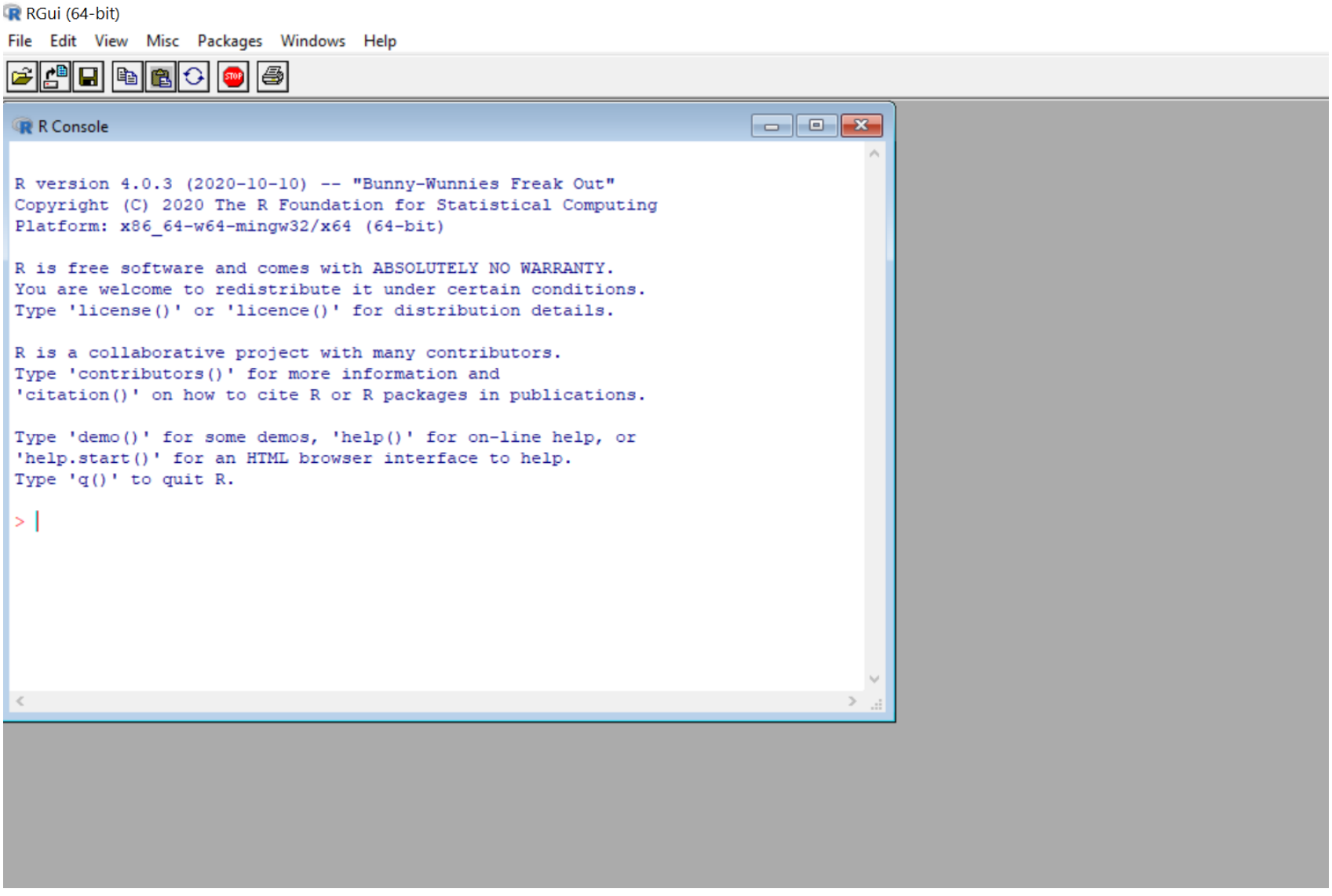
- R and R Studio
- The R syntax
- Figures with ggplot2
- R Markdown/blogdown/bookdown
- #rstats
- Web scrapping with rvest
- OCR with tesseract and daiR
- Future R sessions


AS SEEN BY USERS OF ...
STATA
R
sas
python
SPSS
STATA
R
sas
python
SPSS
My Homework

Base R

- R is the name of a widely-used, open source programming language (so-called "Base R")
- R was originally developed 26 years ago by Ross Ihaka and Robert Gentleman. It is now developed by the R Foundation

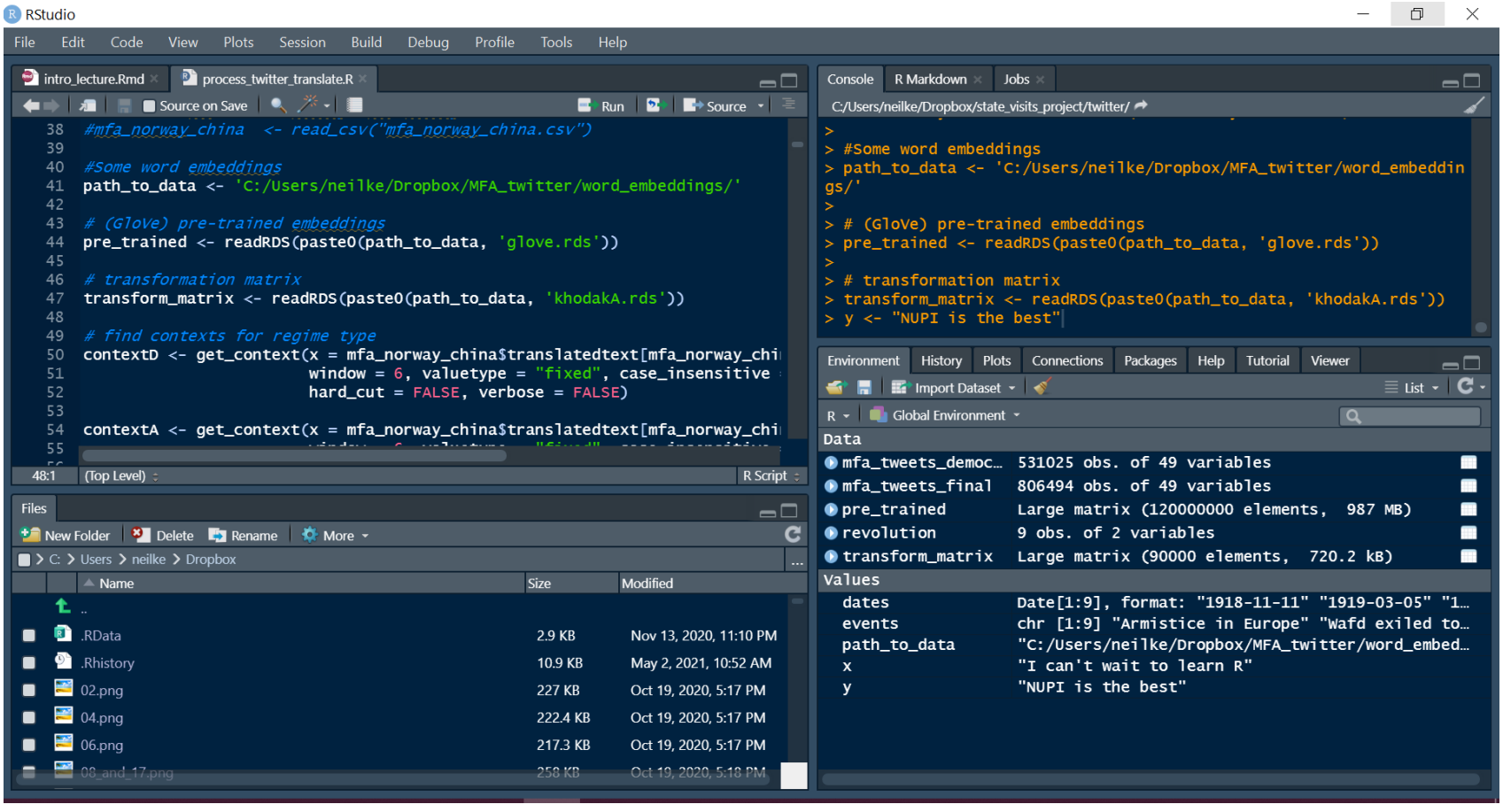
This is what Base R looks like . . . very 1990s



R Studio

- RStudio is an IDE ("Integrated Development Environment"), i.e., an interface for working in R
- RStudio is developed by RStudio Inc., a commercial company
- The free version of R Studio is more than adequate for our needs

This is what R Studio looks like . . . fancy!



Exercise

You can download both R and R Studio on Windows and Mac easily and for free

- Install R for Mac from here: https://cran.r-project.org/bin/macosx/. Install R for Windows from here: https://cran.r-project.org/bin/windows/base/
- Download RStudio for Windows or Mac from here: https://rstudio.com/ products/rstudio/download/, choosing the Free version: this is what most people use and is more than enough for our needs

Check the Teams chat for the URLs

Learning a programming language

The R syntax

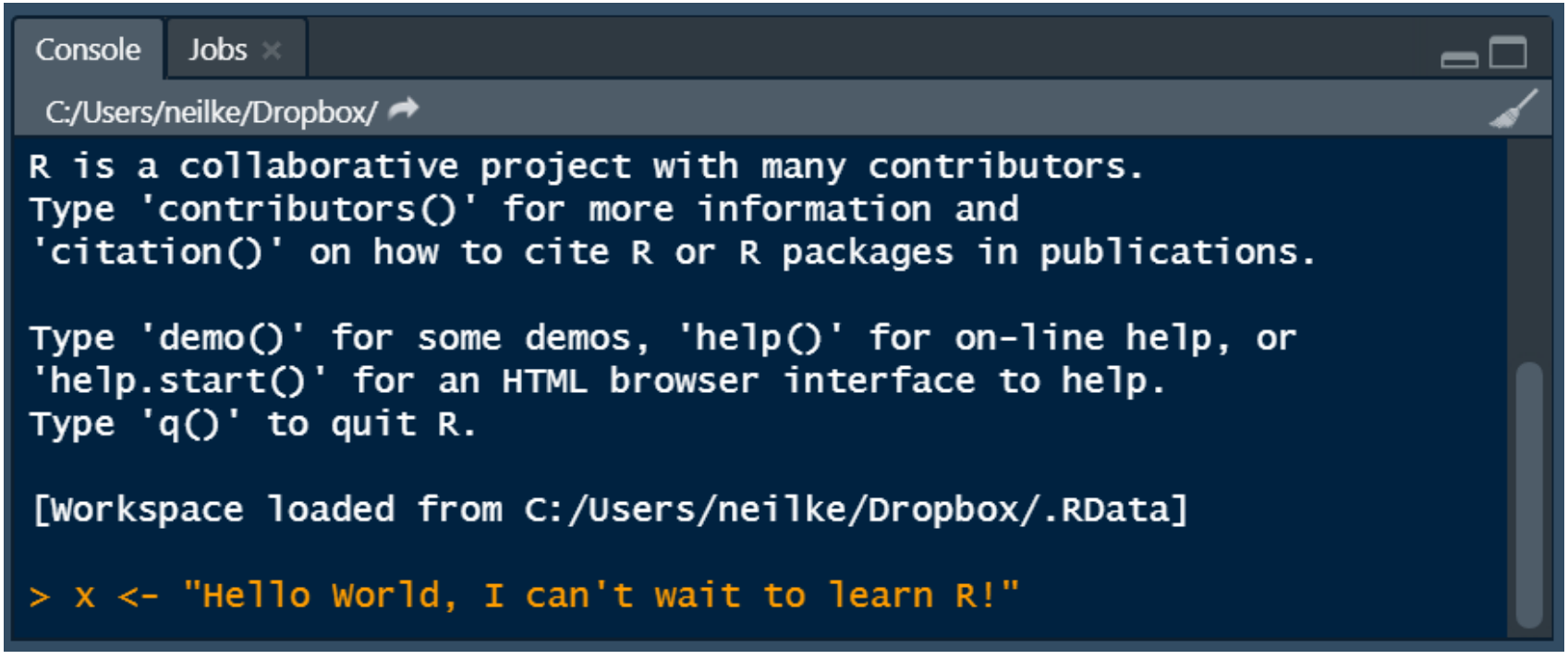
```
x <- "Hello World, I can't wait to learn R!"
```

The R syntax

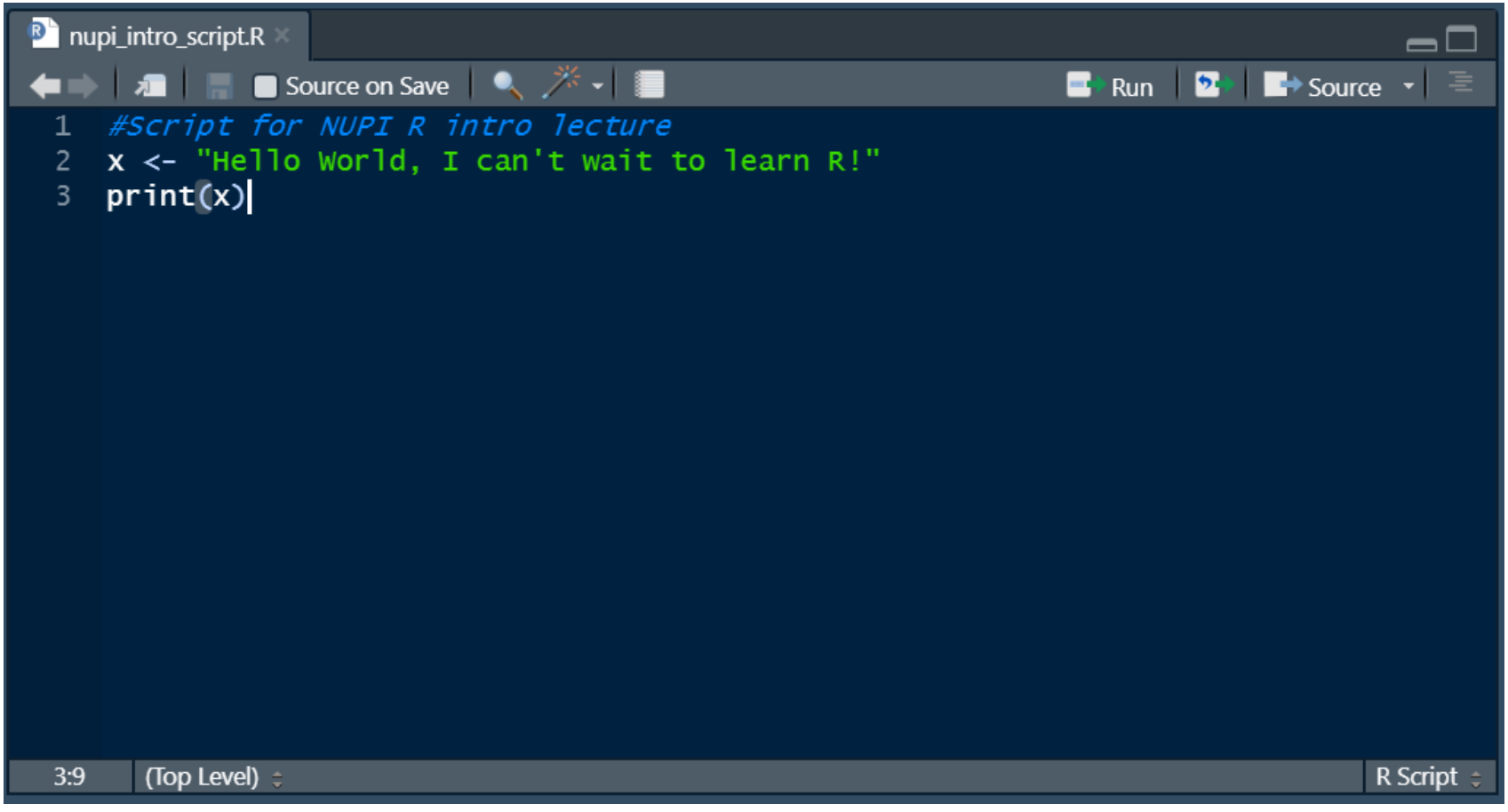
```
print(x)
```

```
## [1] "Hello World, I can't wait to learn R!"
```

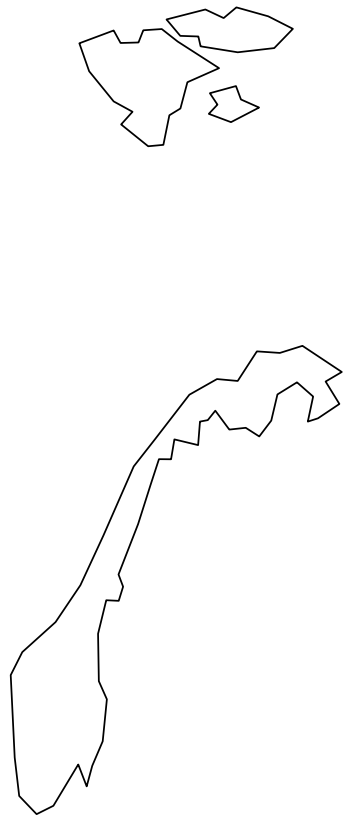
Code can be entered into the command line in the console one line at a time . . .



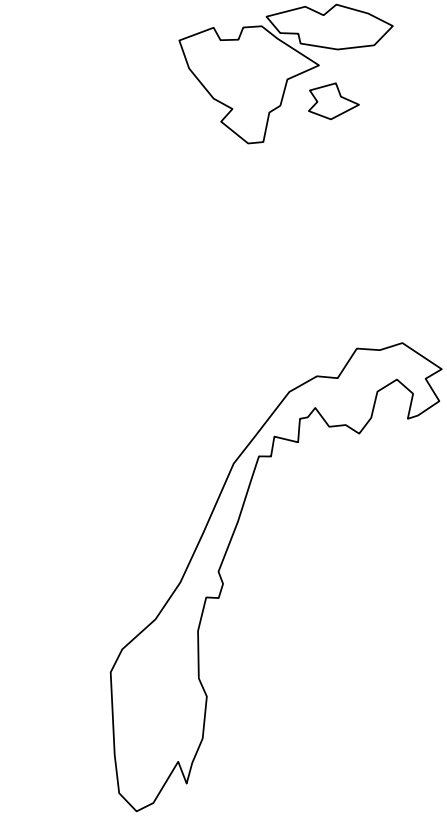
. . . or we can write scripts that execute multiple lines simultaneously



```r
#install.packages("rworldmap")
library(rworldmap)
worldmap <- getMap()
plot(worldmap[which(worldmap$SOVEREIGNT=="Norway"),])
```



```
plot(worldmap[which(worldmap$SOVEREIGNT=="Norway"),],
main = x)
```

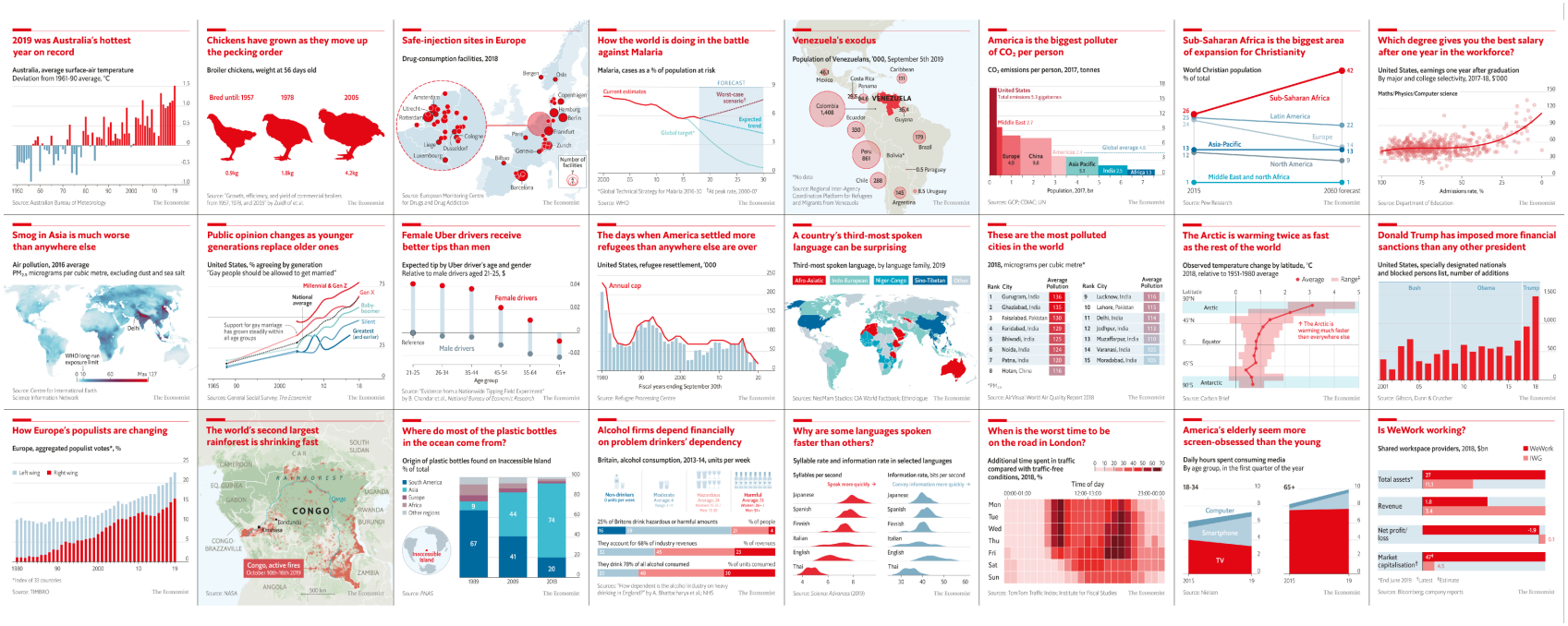


Beautiful graphics

ggplot2

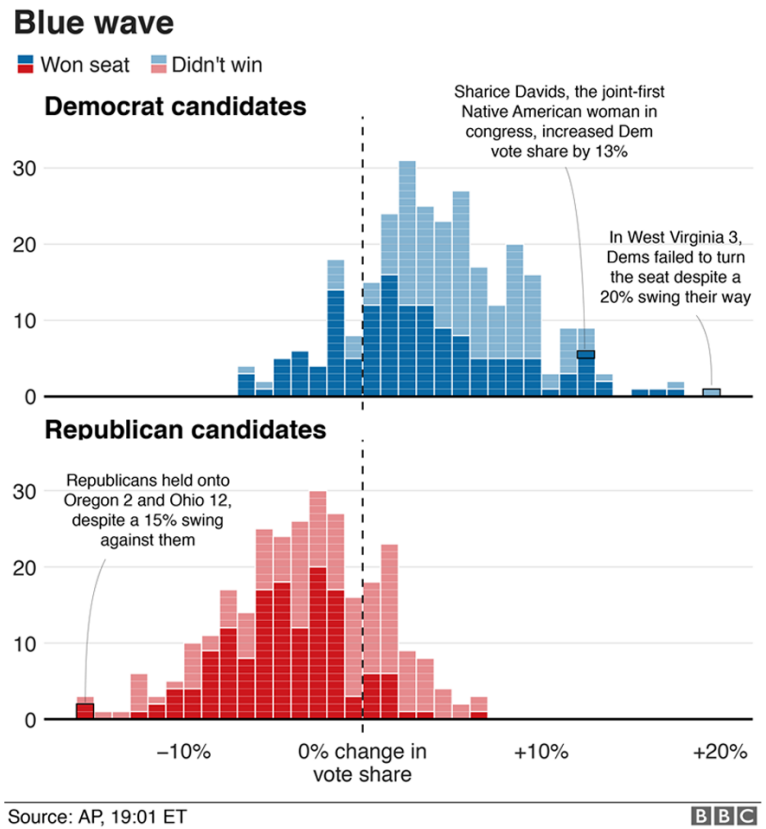
- ggplot2 is a powerful package that allows you to make highly-customizable, publication quality figures
- Many major news organizations make their graphics using ggplot2

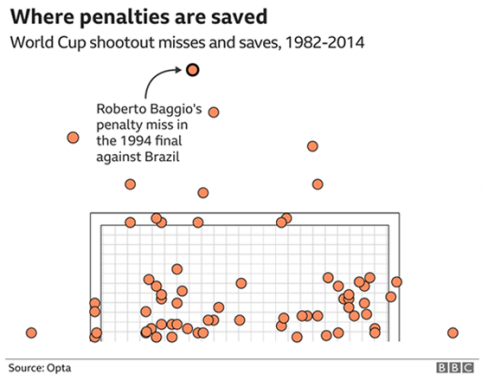
The Economist uses ggplot2 for dataviz

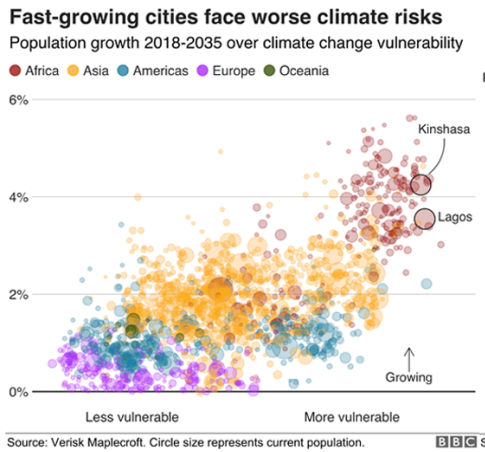


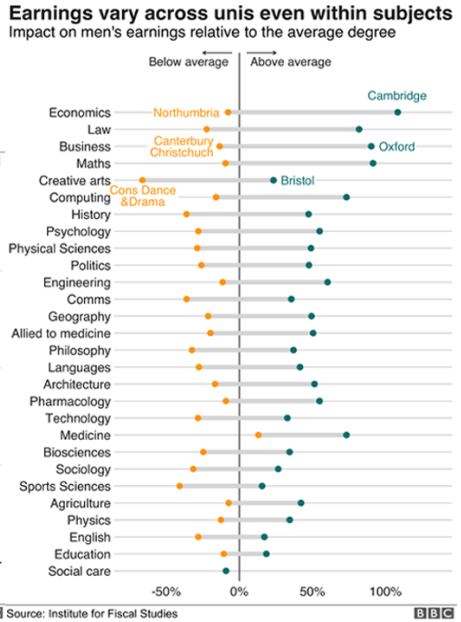
Recreate a graph from *The Economist* in ggplot2: http://rstudio-pubs-static.s3.amazonaws.com/284329_c7e660636fec4a42a09eed968dc47f32.html

BBC R Cookbook

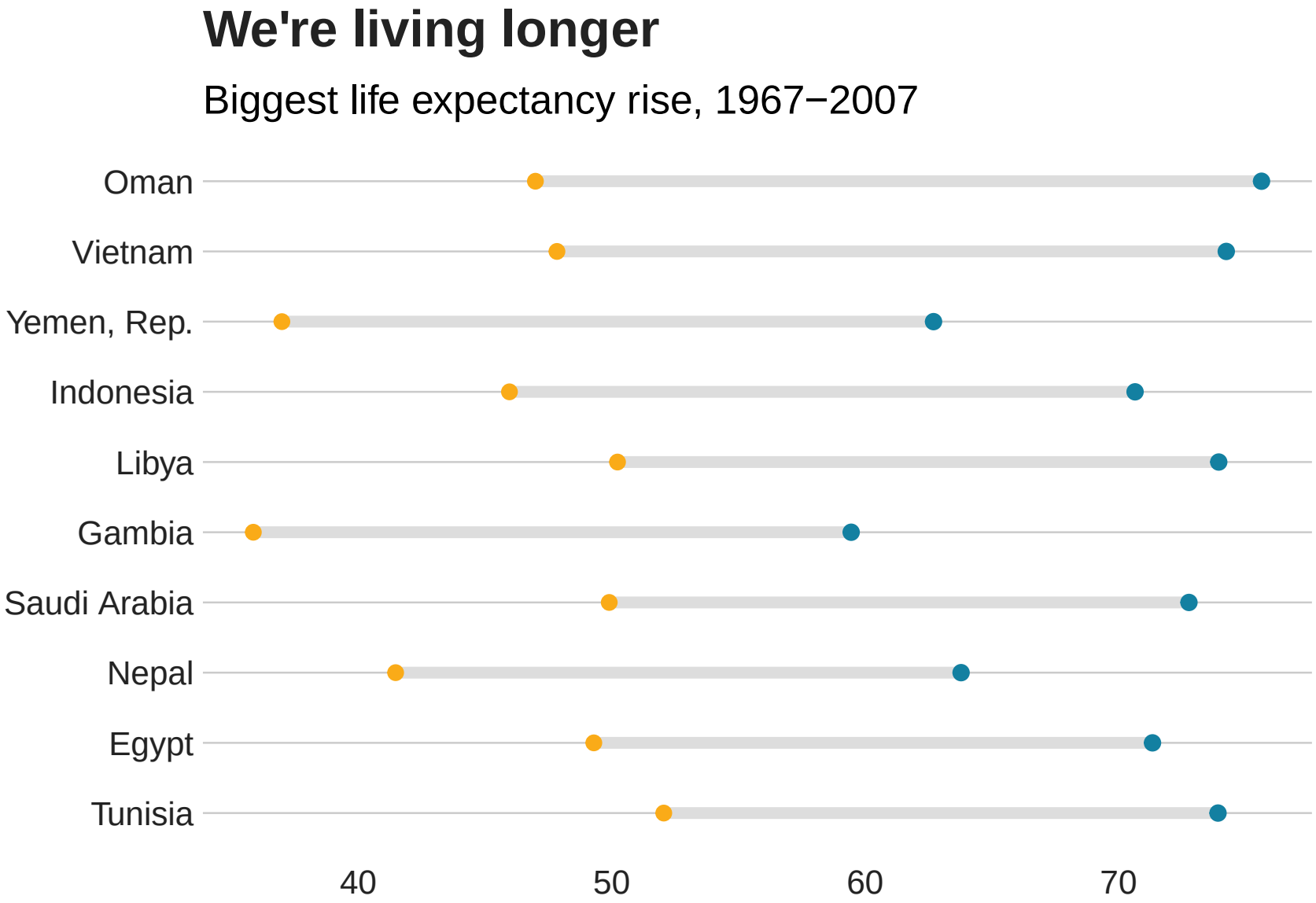








Check out the Cookbook's GitHub page:
https://bbc.github.io/rcookbook/


We're living longer
Biggest life expectancy rise, 1967−2007
Oman
Vietnam
Yemen, Rep.
Indonesia
Libya
Gambia
Saudi Arabia
Nepal
Egypt
Tunisia
40
50
60
70

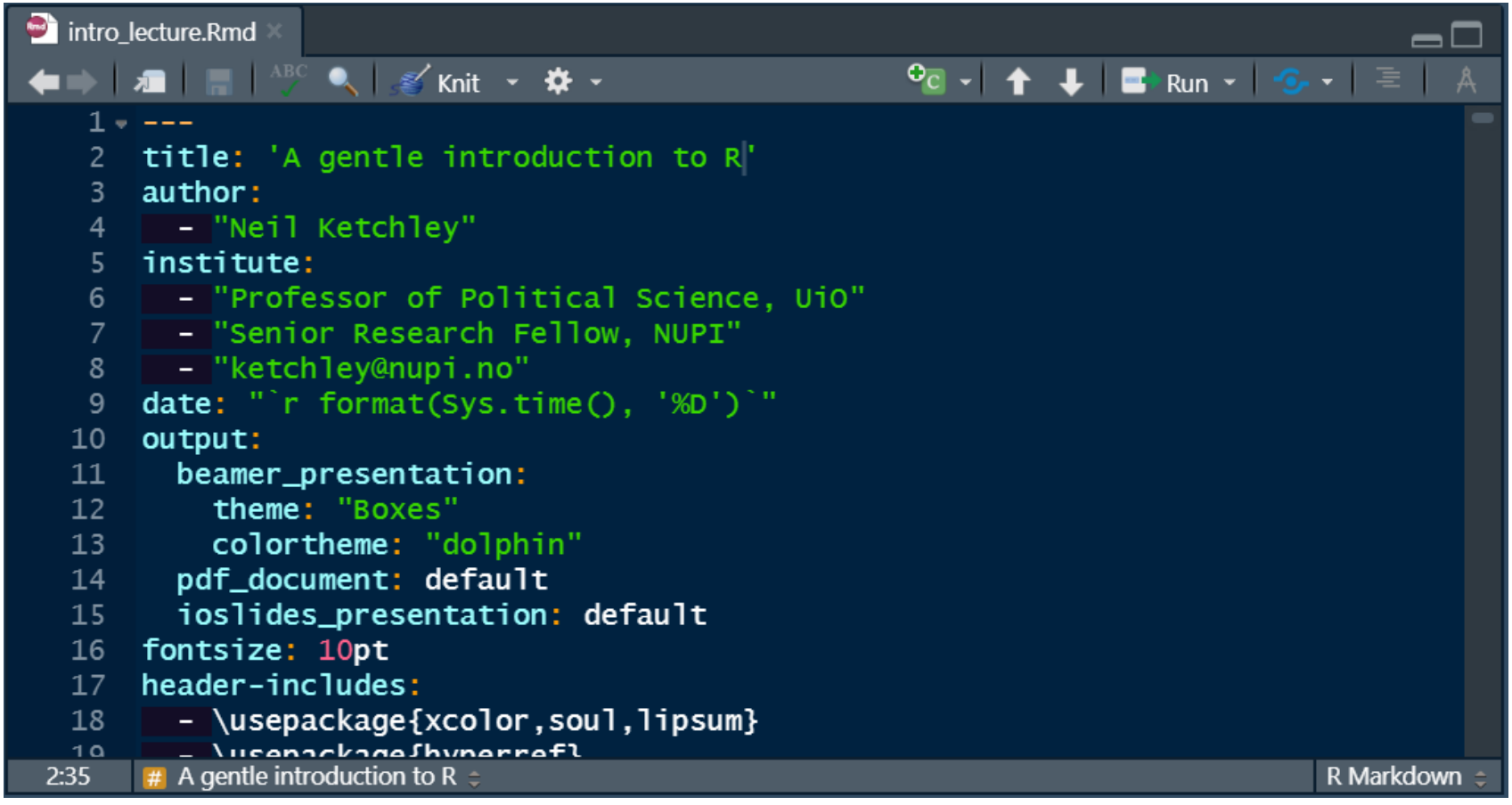
```
library("ggalt")
library("tidyr")
#Prepare data
dumbbell_df <- gapminder %>%
  filter(year == 1967 | year == 2007) %>%
  select(country, year, lifeExp) %>%
  spread(year, lifeExp) %>%
  mutate(gap = `2007` - `1967`) %>%
  arrange(desc(gap)) %>%
  head(10)

#Make plot
ggplot(dumbbell_df, aes(x = `1967`, xend = `2007`,
  y = reorder(country, gap), group = country)) +
  geom_dumbbell(colour = "#dddddd",
                size = 3,
                colour_x = "#FAAB18",
                colour_xend = "#1380A1") +
  bbc_style() +
  labs(title="We're living longer",
       subtitle="Biggest life expectancy rise, 1967-2007")
```

Disseminating reproducible research

R Markdown

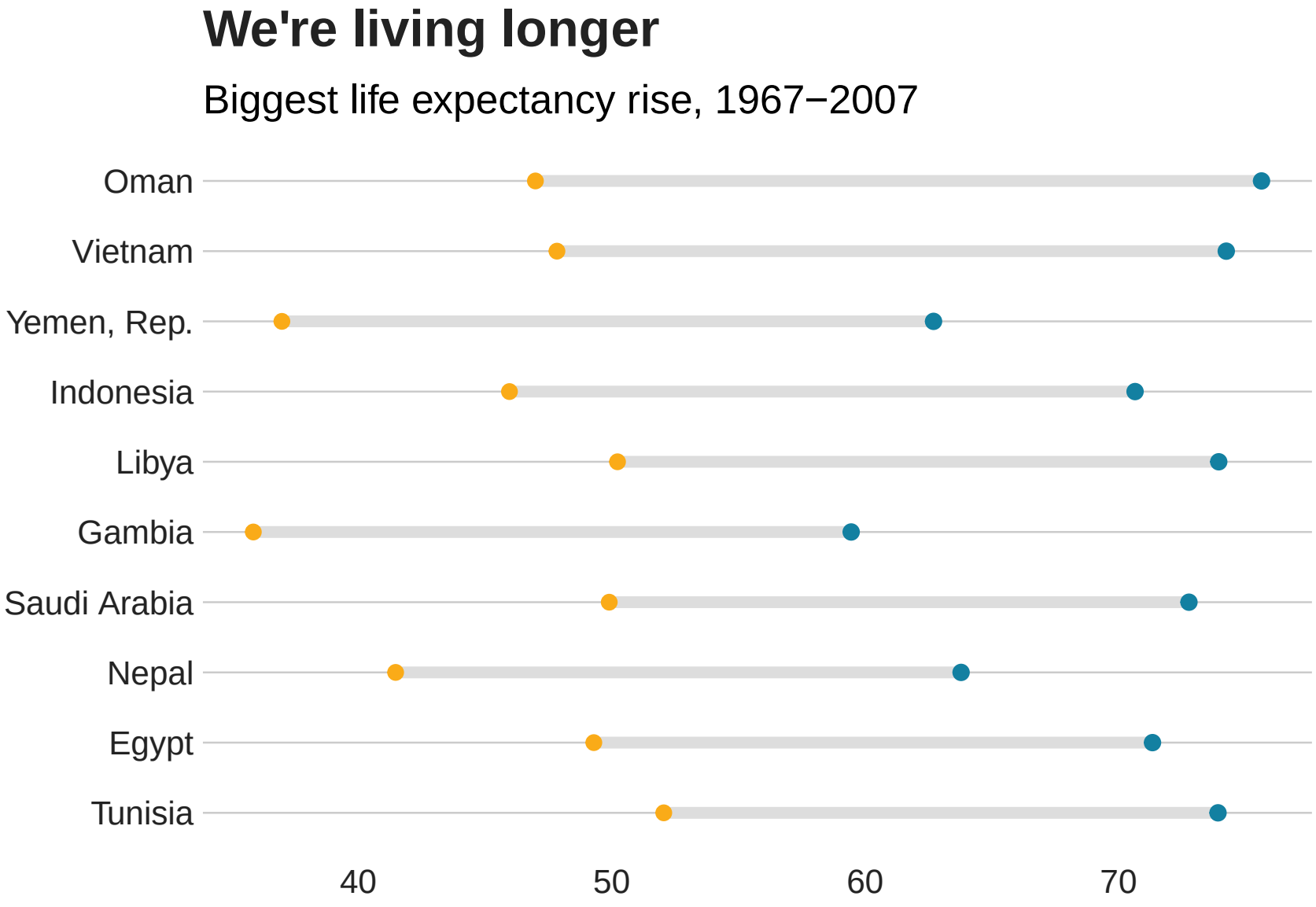
- R markdown can be used to save and execute code
- R Markdown allows you to generate high quality documents that can be shared with an audience

I made these slides in R Markdown!

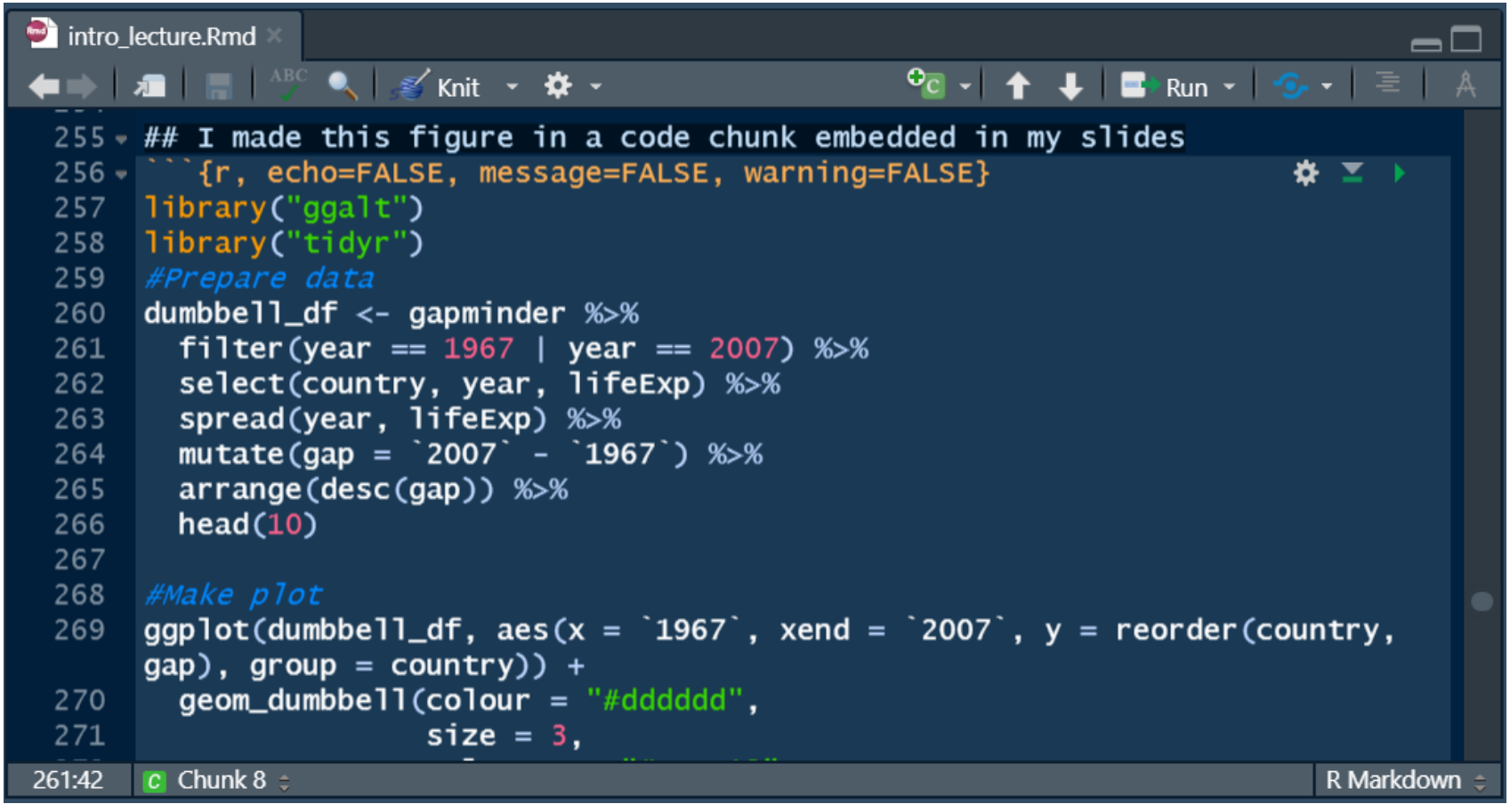


Check out the presentation options in R Markdown:
https://bookdown.org/yihui/rmarkdown/presentations.html

I made this figure in a code chunk embedded in my slides



This means you can update your figures on the fly



We wrote this paper in R Markdown!

Plots, Attacks, and the Measurement of Terrorism*

Thomas Hegghammer† Neil Ketchley‡

March 30, 2021

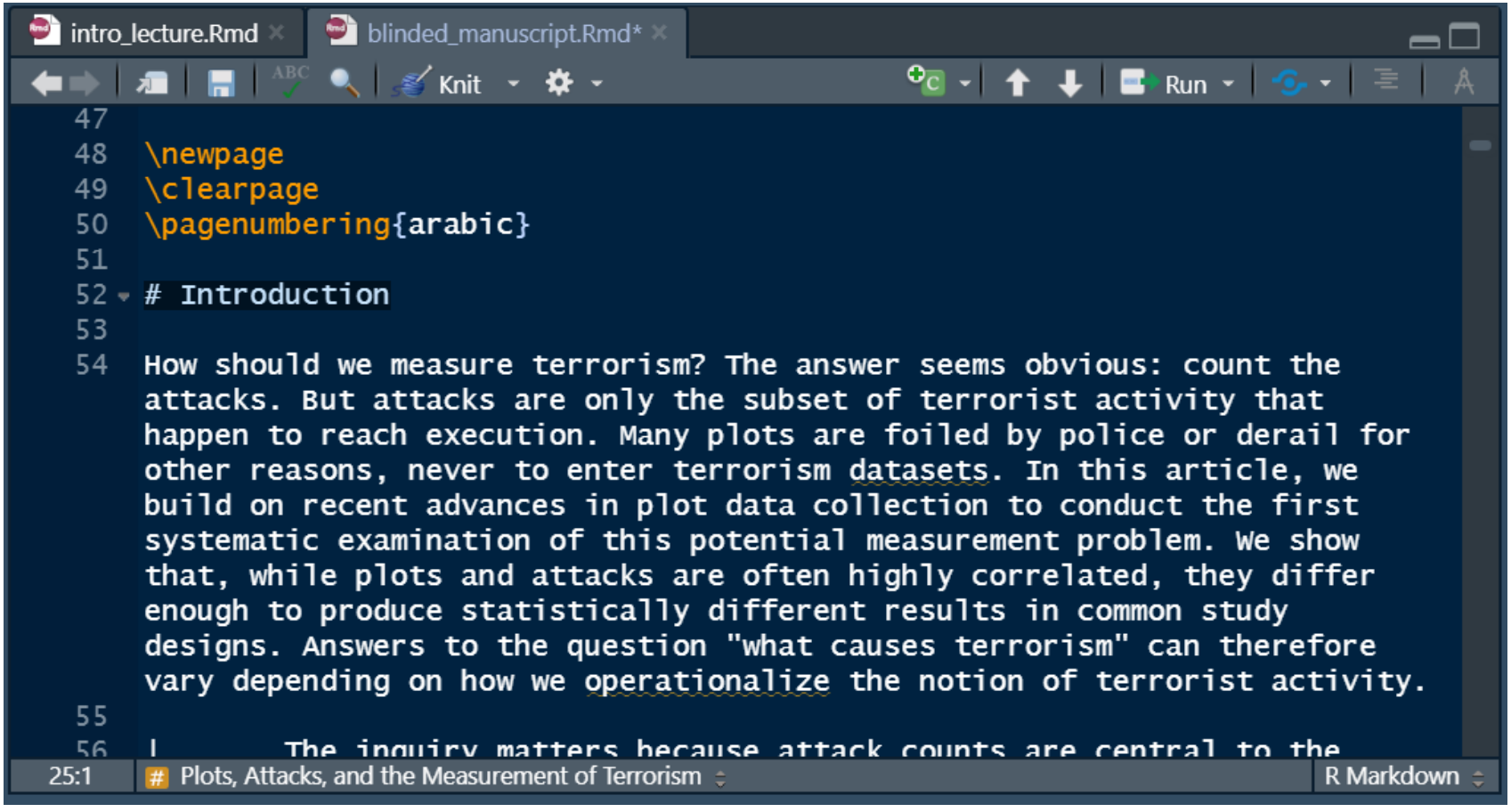
Abstract

Event datasets are central to the study of terrorism. Political scientists typically use terrorist attacks as the dependent variable and test covariates to identify factors that produce terrorism. But attacks are an imperfect measure of terrorist activity because of "plot attrition" — the tendency for plots to derail due to police intervention or other reasons. Building on recent advances in plot data collection, we show that common research designs predicting terrorist incidents produce statistically significantly different results depending on whether incidents are operationalized as plots or attacks. This is after accounting for state security capability, suggesting plot attrition is dynamic and not easily controlled out. Plot data should be incorporated, when available, in future studies on the causes of terrorism.

Keywords: Terrorism, data, measurement

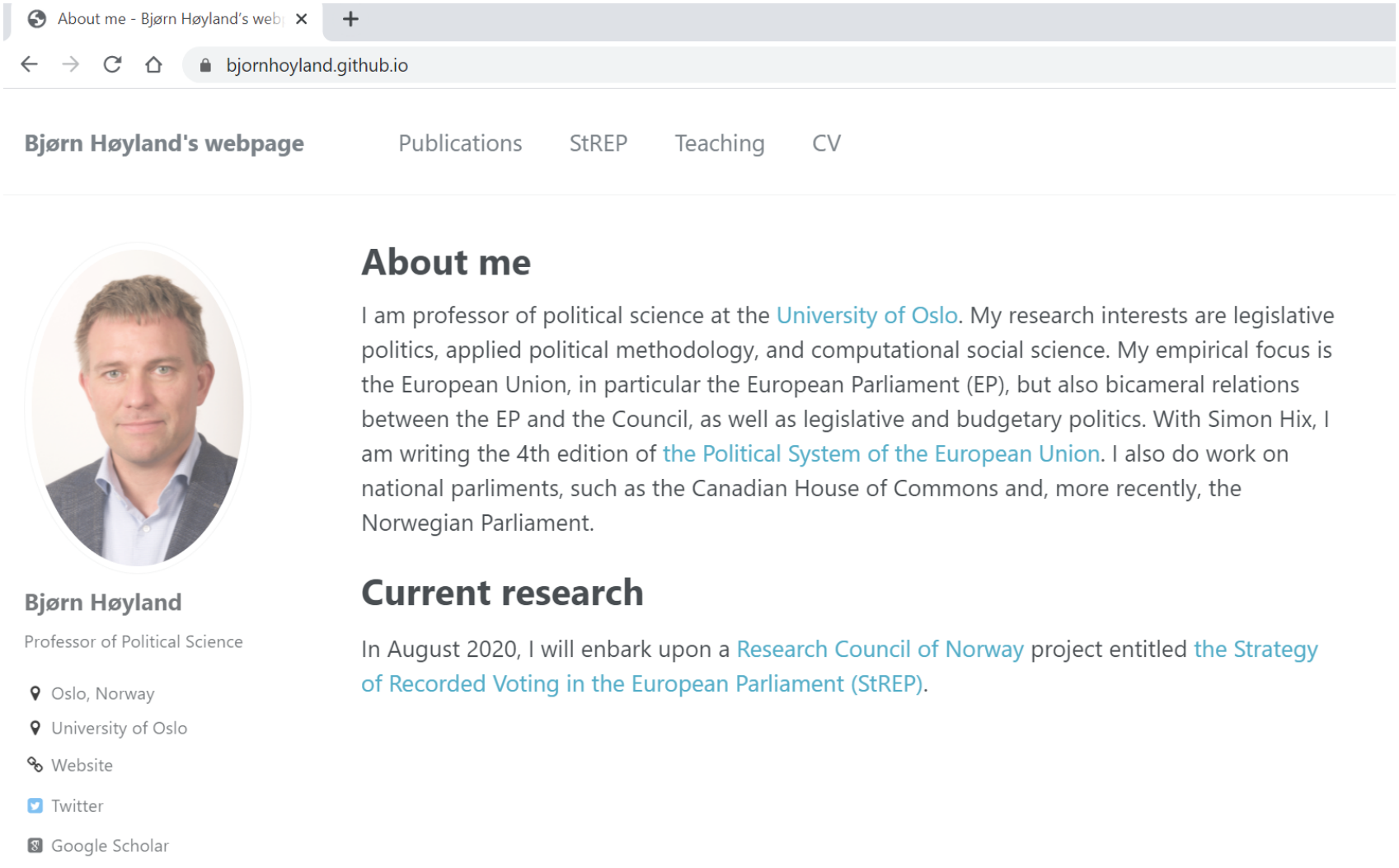
Check out the paper: https://osf.io/preprints/socarxiv/t72yj/

We wrote this paper in R Markdown!



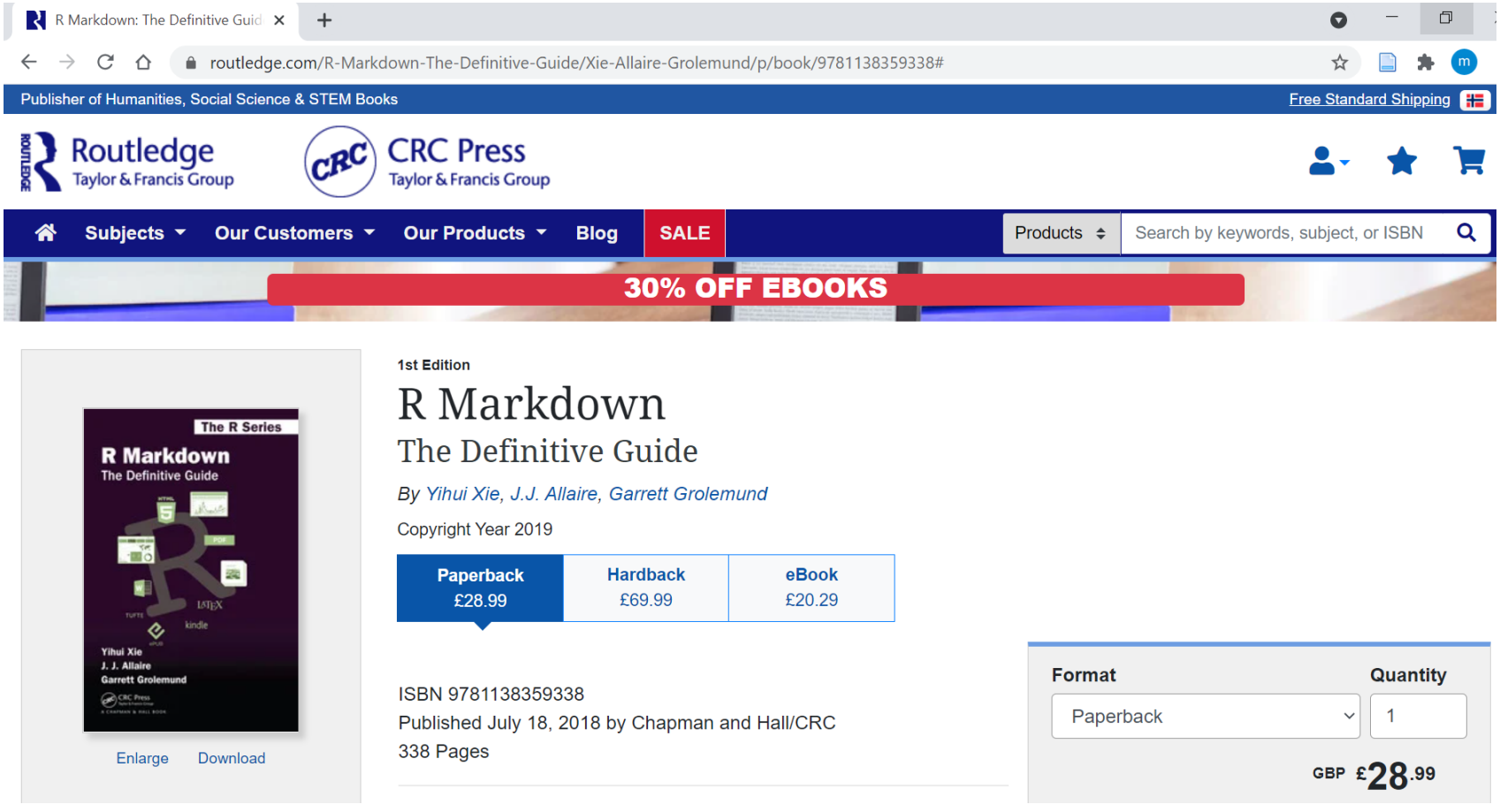
Check out the article options in R Markdown:
https://bookdown.org/yihui/rmarkdown/journals.html

My colleagues make their websites in R Blogdown!

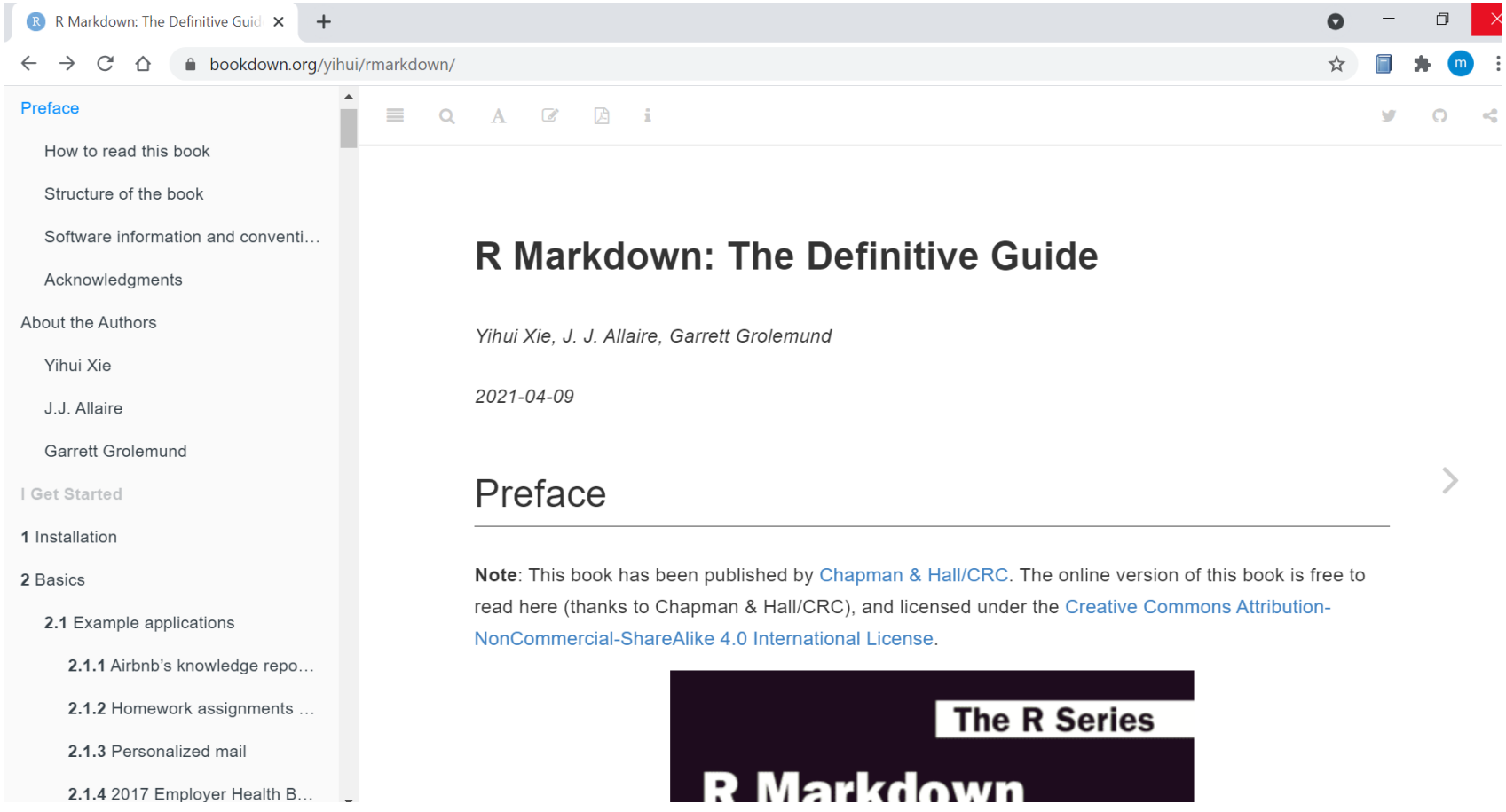


Check out Blogdown: https://bookdown.org/yihui/blogdown/

Books about R are made in R Markdown . . .



. . . and are available online for free via R Bookdown



Check out the Bookdown about R Markdown:
https://bookdown.org/yihui/rmarkdown/

#rstats

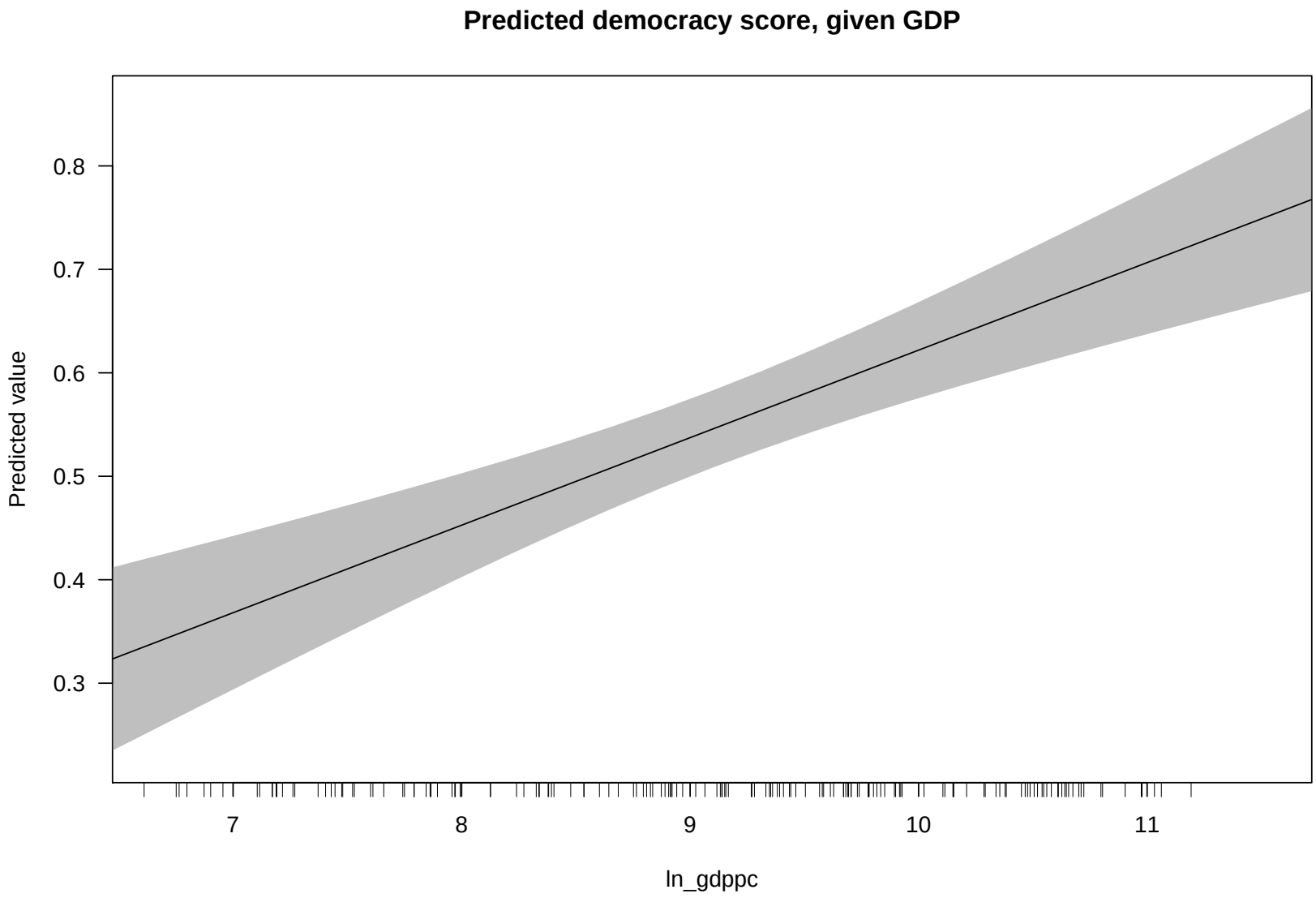
```
library(stargazer)
mod <- lm(v2x_polyarchy ~ ln_gdppc, data = dat)
stargazer(mod, header = FALSE, style = "apsr",
          title = "How democracy increases with GDP")
```

Table 1: How democracy increases with GDP

	v2x_polyarchy
ln_gdppc	0.085^{***}
	(0.016)
Constant	-0.224
	(0.143)
N	162
R^2	0.155
Adjusted R^2	0.150
Residual Std. Error	0.242 (df = 160)
F Statistic	29.379^{***} (df = 1; 160)

*p $<$ .1; **p $<$ .05; ***p $<$ .01

```
library(margins)
cplot(mod, "ln_gdppc", what = "prediction",
      main = "Predicted democracy score, given GDP")
```



Web scrapping with rvest

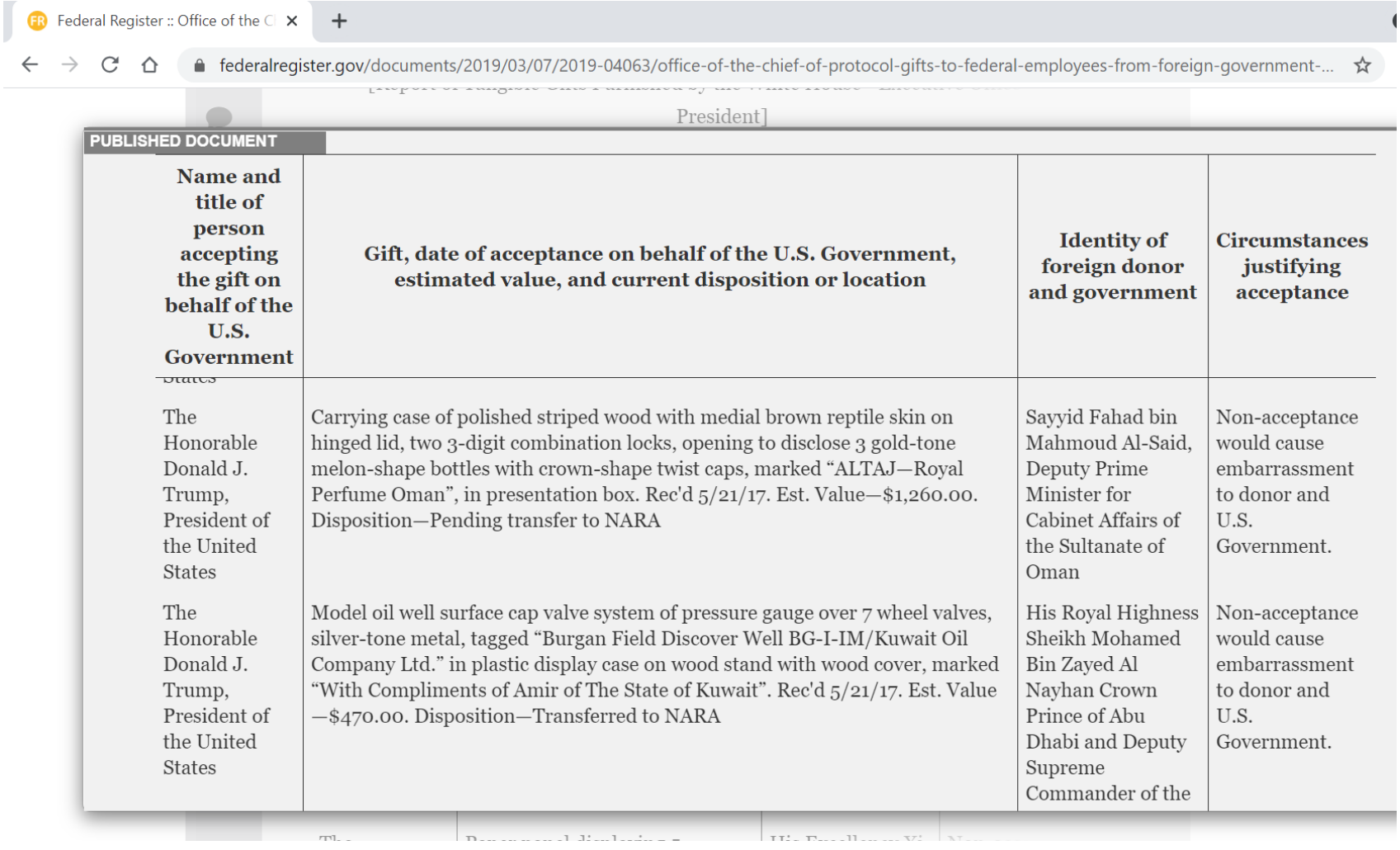
rvest

- rvest helps you scrape (or harvest) data from web pages
- It does so by using the elements contained within an HTML page

Diplomatic gifts to the U.S.

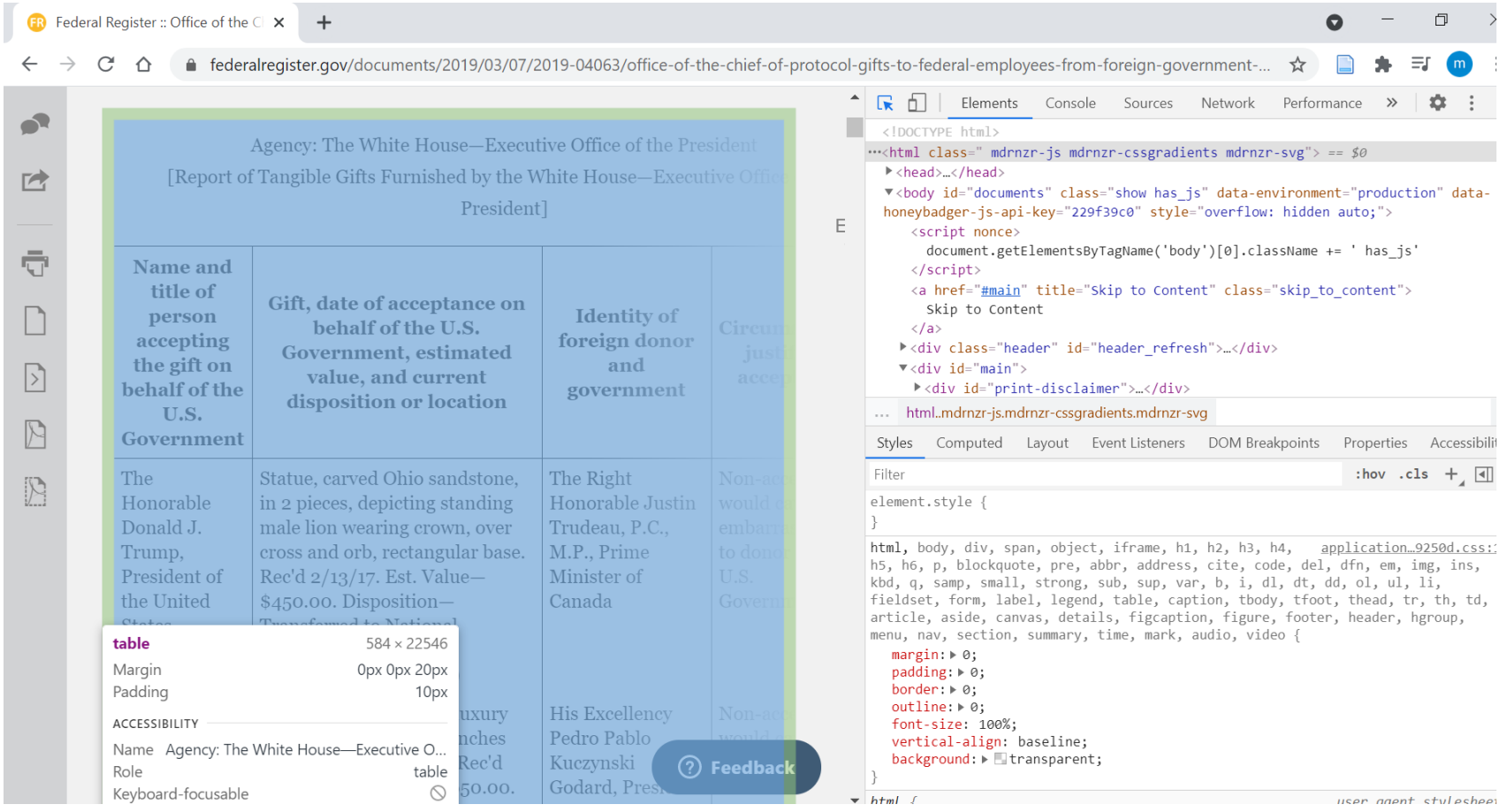
Source: https://www.federalregister.gov/documents/2019/03/07/2019-04063/office-of-the-chief-of-protocol-gifts-to-federal-employees-from-foreign-government-sources-reported

Web pages often contain tables and other data that we want to analyze



Name and title of person accepting the gift on behalf of the U.S. Government	Gift, date of acceptance on behalf of the U.S. Government, estimated value, and current disposition or location	Identity of foreign donor and government	Circumstances justifying acceptance
The Honorable Donald J. Trump, President of the United States	Carrying case of polished striped wood with medial brown reptile skin on hinged lid, two 3-digit combination locks, opening to disclose 3 gold-tone melon-shape bottles with crown-shape twist caps, marked "ALTAJ—Royal Perfume Oman", in presentation box. Rec'd 5/21/17. Est. Value—$1,260.00. Disposition—Pending transfer to NARA	Sayyid Fahad bin Mahmoud Al-Said, Deputy Prime Minister for Cabinet Affairs of the Sultanate of Oman	Non-acceptance would cause embarrassment to donor and U.S. Government.
The Honorable Donald J. Trump, President of the United States	Model oil well surface cap valve system of pressure gauge over 7 wheel valves, silver-tone metal, tagged "Burgan Field Discover Well BG-I-IM/Kuwait Oil Company Ltd." in plastic display case on wood stand with wood cover, marked "With Compliments of Amir of The State of Kuwait". Rec'd 5/21/17. Est. Value —$470.00. Disposition—Transferred to NARA	His Royal Highness Sheikh Mohamed Bin Zayed Al Nayhan Crown Prince of Abu Dhabi and Deputy Supreme Commander of the	Non-acceptance would cause embarrassment to donor and U.S. Government.

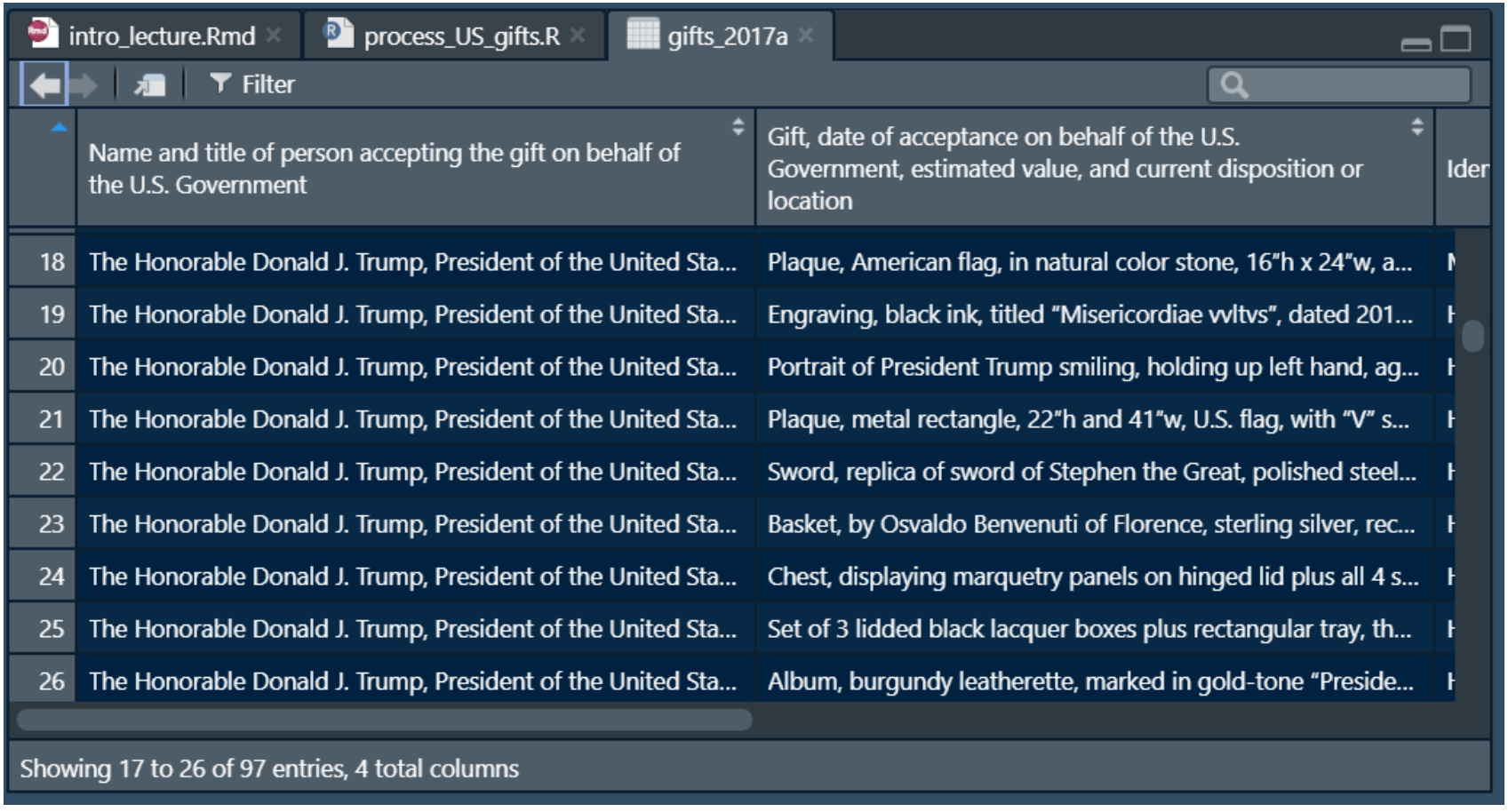
We can select the relevant xpath using the “inspect” option



Scrape the page using rvest

```
library("rvest")
url <- "https://www.federalregister.gov/documents/2019/03/07/201
gifts_2017a <- url %>%
  html() %>%
  html_nodes(xpath=
  '//*[@id="fulltext_content_area"]/div[4]/table') %>%
  html_table()
gifts_2017a <- gifts_2017a[[1]]
```

We now have a diplomatic gifts dataset!



	Name and title of person accepting the gift on behalf of the U.S. Government	Gift, date of acceptance on behalf of the U.S. Government, estimated value, and current disposition or location	Ider
18	The Honorable Donald J. Trump, President of the United Sta...	Plaque, American flag, in natural color stone, 16"h x 24"w, a...	M
19	The Honorable Donald J. Trump, President of the United Sta...	Engraving, black ink, titled "Misericordiae vvltvs", dated 201...	H
20	The Honorable Donald J. Trump, President of the United Sta...	Portrait of President Trump smiling, holding up left hand, ag...	H
21	The Honorable Donald J. Trump, President of the United Sta...	Plaque, metal rectangle, 22"h and 41"w, U.S. flag, with "V" s...	H
22	The Honorable Donald J. Trump, President of the United Sta...	Sword, replica of sword of Stephen the Great, polished steel...	H
23	The Honorable Donald J. Trump, President of the United Sta...	Basket, by Osvaldo Benvenuti of Florence, sterling silver, rec...	H
24	The Honorable Donald J. Trump, President of the United Sta...	Chest, displaying marquetry panels on hinged lid plus all 4 s...	H
25	The Honorable Donald J. Trump, President of the United Sta...	Set of 3 lidded black lacquer boxes plus rectangular tray, th...	H
26	The Honorable Donald J. Trump, President of the United Sta...	Album, burgundy leatherette, marked in gold-tone "Preside...	H

Showing 17 to 26 of 97 entries, 4 total columns

OCR

tesseract

- We often want to extract text from an image
- R has open source OCR libraries to do this, e.g. tesseract. These work best for images with high contrast, little noise and horizontal text
- You can also use R to interact with paid-for cloud services, e.g. Google Document AI, which are much more powerful

Colonial diplomatic telegrams

Sir M. Cheetham to Earl Curzon.—(Received March 19.)

(No. 406. Urgent.)
(Telegraphic.) *Cairo, March 18, 1919.*

FOLLOWING is summary of events reported during the twenty-four hours ending noon, 18th March, 1919 :—

Between Cairo, Alexandria, and Port Said various stations burnt signal-boxes destroyed, and rails and sleepers removed along the main and subsidiary lines.

One train arrived at Cairo from Port Said after journey of twenty-six hours, preceded by military train repairing the line. No communications with Upper Egypt beyond Wasta, where trouble serious. Many other stations destroyed. Similar damage further south reported by telegram viâ Port Sudan ; natives constantly engaged in destroying lines in numerous districts until dispersed by fire.

Telegraph wires and poles being destroyed in all directions.

There is no communication other than by wireless telegraph and aeroplane between Cairo and the provinces.

At Cairo a peaceful demonstration, in which some thousands took part, headed by university intellectuals, took place yesterday without incident, except where once interfered with by party of soldiers, one native casualty resulting.

Demonstration this morning at Boulac was dispersed with several casualties.

Further demonstration in Cairo was ordered to disperse by military authorities and did so.

At Alexandria the Khedivial mail line employees joined the students. Trouble ensued ; several casualties caused by troops ; dock gates closed.

Various grain, sugar, and oil stores threatened, but precautions taken. Attacks were made on sugar factories at Cairo, but assailants beaten off.

OCR with tesseract

```
library(tesseract)
eng <- tesseract("eng")
text <- tesseract::ocr("C:/Users/neilke/Dropbox/NUPI/NUPI_R/img1
      , engine = eng)
```

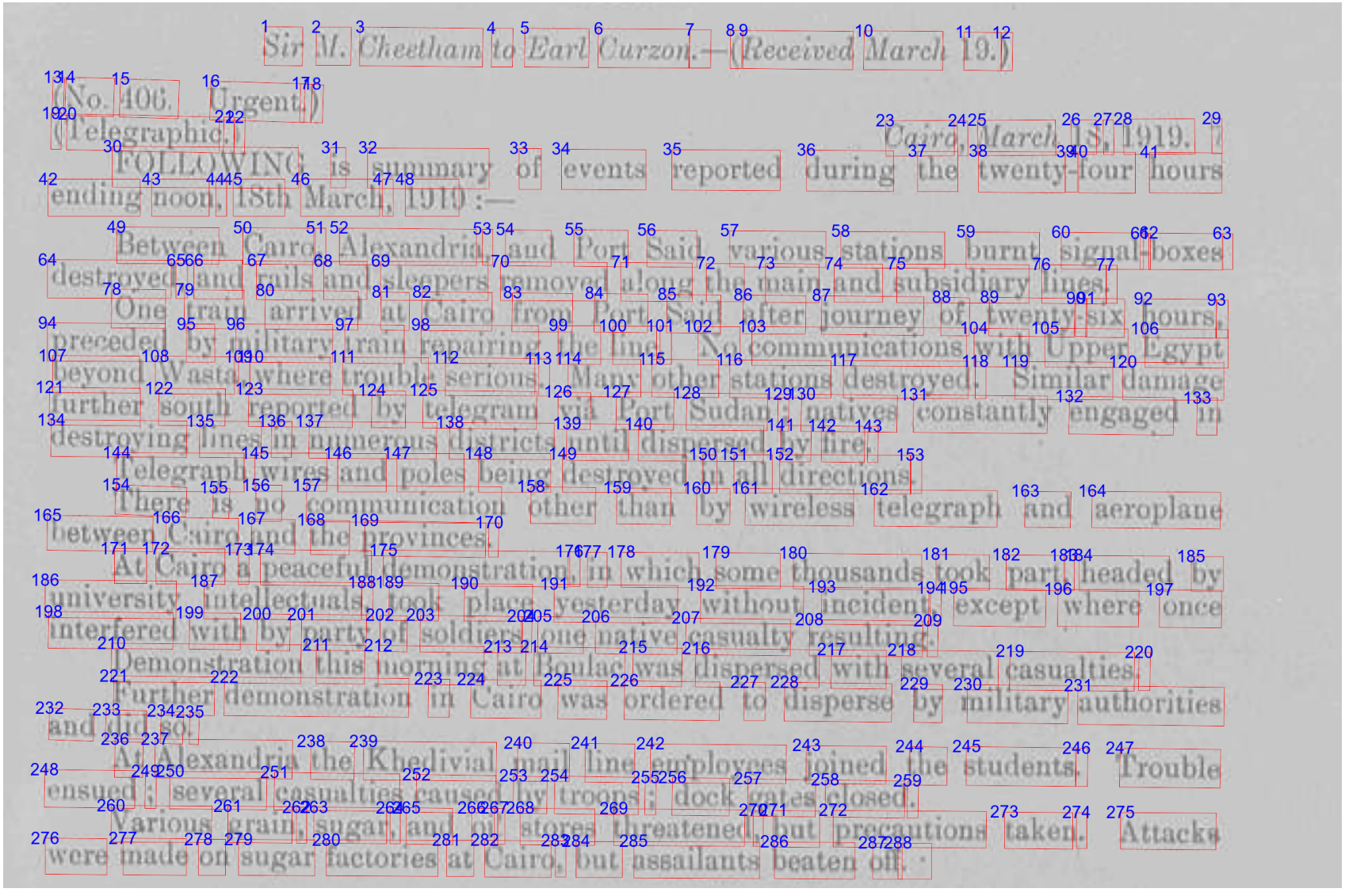
Sir M. Cheetham to art Curzon.—(Received **Mareh** 19.) (No. 406, Urgent.) (Telegraphic.) Cairo, **Mareh** 18, 1919. 7 FOLLOWING is summary of events reported during the twenty-four hours ending noon, 18th March, 1010/:— Between Cairo, Alexandria, **aud** Port Said various stations burnt signal-boxes **destrayed**, and rails and sleepers removed **wong** the main and subsidiary lines. One train arrived at Cairo ftom Port Said after journey of twenty-six hours, **preeeded** by military wain repairing the line, No **communieations** with Upper Egypt beyond Wasta, where trouble serious. Many other stations destroyed. Similar damage further south reported by telegram **yid** Port Sudan; natives constantly engaged in destroying lines in numerous districts until dispersed by fire. **F** 'Telegraph wires and poles being **destroved** in all directions. There is no communication other than by wireless telegraph and aeroplane **hetween** Cairo and the provinces. At Cairo a **aiaet** demonstration, in which some thousands took part, headed by university intellectuals, took place yesterday without incident, except where once interfered with by party of soldiers, one native casualty resulting. **i** Demonstration this morning at **Boulae** was dispersed with several casualties. Further demonstration in Cairo was ordered to disperse by military authorities and did so. At Alexandria the Khedivial mail line employees joined the students, 'Trouble ensued ; several casualties caused by troops; dock gates closed. Various grain, sugar, and **oi)** stores threatened, but precautions taken. Attacks were made on sugar factories at Cairo, but assailants beaten off,

Accessing Google Document AI using daiR

```
#library(dair)
#library(googleCloudStorageR)
#my_project_id2 <- "ocr-1919-telegrams"
#gcs_list_buckets(my_project_id2)
#gcs_create_bucket("ocr_1919_telegrams", my_project_id, location
#pdfs <- dir_ls(glob = "*.pdf")
#map(pdfs, ~ gcs_upload(.x, name = paste0("historical/", .x)))

#content <- gcs_list_objects()
#our_files <- grep("^historical/mansoura_telegram", content$name

#response <- dai_process(our_files, "processed")

#jsons <- grep("*.json$", contents$name, value = TRUE)
#map(jsons, ~ gcs_get_object(.x, saveToDisk = gsub("/", "_", .x)

#text1 <- get_text("processed_mansoura_telegram.json")
#draw_tokens("mansoura_telegram.pdf", "processed_mansoura_telegr
```

Check out the daiR package: https://dair.info/

OCR with daiR



OCR results with daiR

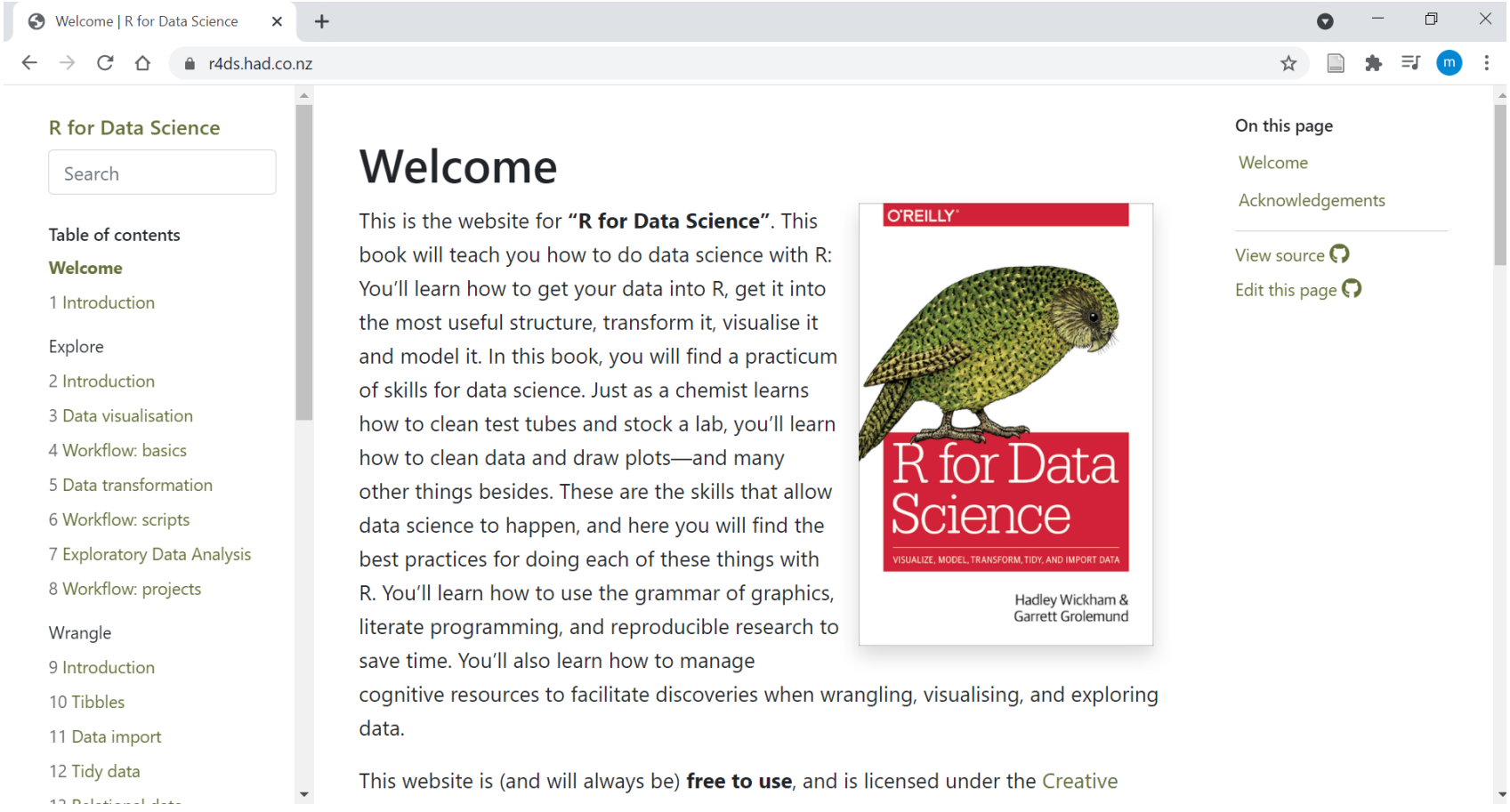
Sir M. Cheetham to Earl Curzon.-(Received March 19.) (No. 406. Urgent.) (Telegraphic.) Cairo, March 18, 1919. 7 FOLLOWING is summary of events reported during the twenty-four hours ending noon, 18th March, 1919- Between Cairo, Alexandria, and Port Said various stations burnt signal-boxes: destroyed, and rails and sleepers removed along the main and subsidiary lines. One train arrived at Cairo from Port Said after journey of twenty-six hours, preceded by military train repairing the line. No communications with Upper Egypt beyond Wasta, where trouble serious. Many other stations destroyed. Similar damage further south reported by telegram **viâ** Port Sudan; natives constantly engaged in destroying lines in numerous districts until dispersed by fire. Telegraph wires and poles being destroyed in all directions. There is no communication other than by wireless telegraph and aeroplane between Cairo and the provinces. At Cairo a peaceful demonstration, in which some thousands took part, headed by university intellectuals, took place yesterday without incident, except where once interfered with by party of soldiers, one native casualty resulting. Demonstration this morning at Boulac was dispersed with several casualties. Further demonstration in Cairo was ordered to disperse by military authorities and did so. At Alexandria the Khedivial mail line employees joined the students. Trouble ensued; several casualties caused by troops; dock gates closed. Various grain, sugar, and oil stores threatened, but precautions taken. Attacks were made on sugar factories at Cairo, but assailants beaten off.

daiR can handle over 60 languages

Upcoming NUPI research design seminars

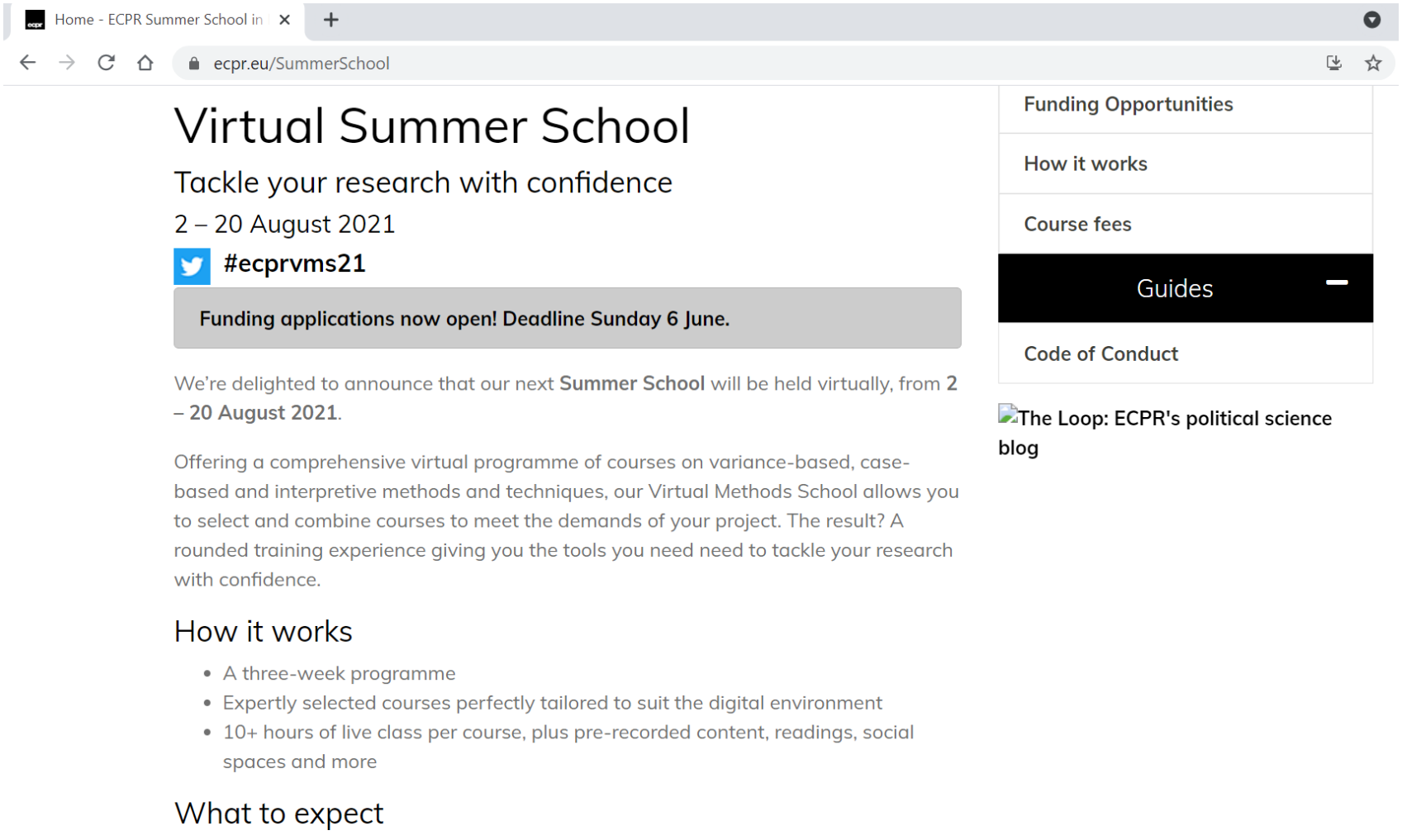
- Web scrapping - 3 June
- Social network analysis - 10 June
- Text analysis - 24 June
- R lab sessions - after the summer

Further reading: R for Data Science



Read it for free: https://r4ds.had.co.nz/

ECPR Summer and Winter Schools



Register here: https://ecpr.eu/SummerSchool

Try out the exercises yourself

Download the R Markdown file and the data needed to run the code chunks from this GitHub repository:
https://github.com/neilketchley/NUPI

Acknowledgements

- The code for the life expectancy graph comes from the BBC's R Cookbook: https://bbc.github.io/rcookbook/
- The cleaned and trimmed V-Dem data comes from Chris Hanretty's introduction to R: http://chrishanretty.co.uk/conveRt/#1. The full V-Dem data can be downloaded here: https://www.v-dem.net/en/data/data/