

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

Thomas J. Leeper (@thosleeper) · 7 Jul 2019

Raise your hand if you learned to code by copy-pasting stuff from the internet and changing it until it broke. 🤚

1.09 PM - 7 Jul 2019

135 Retweets 2,096 Likes

46 135 2.1K

Thomas J. Leeper (@thosleeper) · 16h

Beautiful analogy from @AnitaBlanchard:

Anita Blanchard (@AnitaBlanchard) · 16h

Replies to @drschweitzer

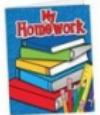
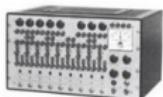
I consider this mode of programming to be like baking sour dough bread:
you are just using a little starter!

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 ...

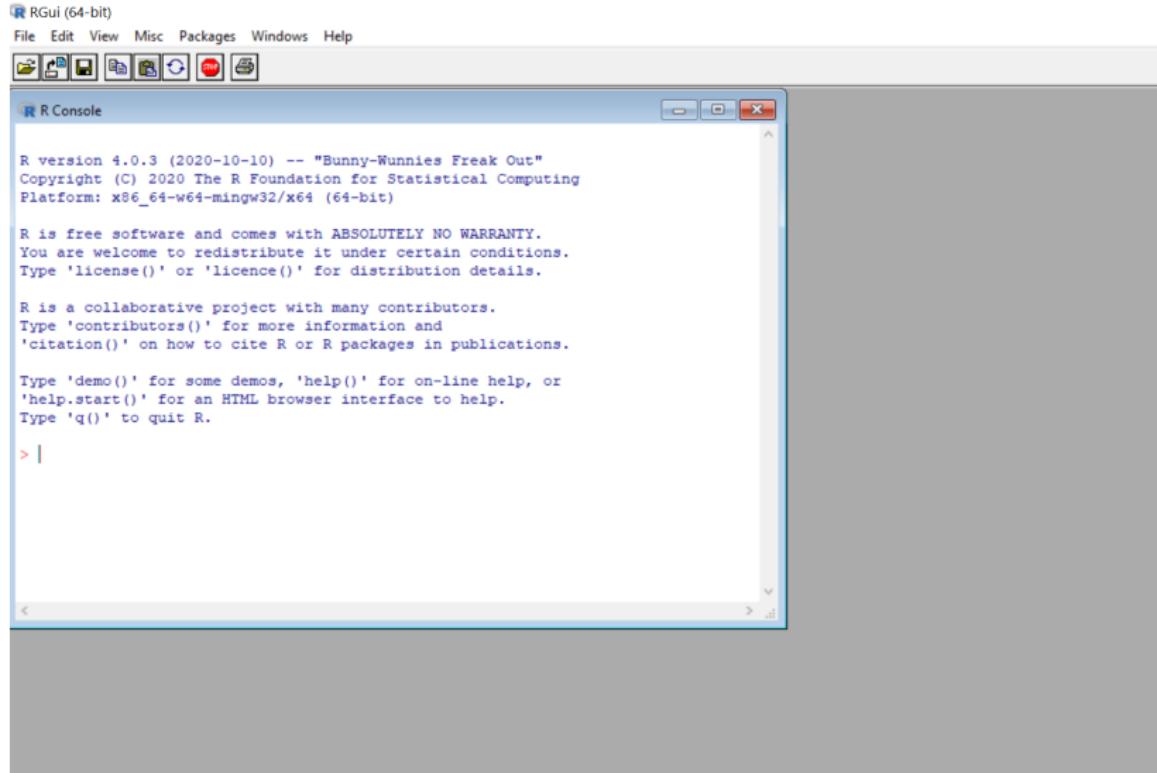


Base R



- ▶ 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!

The screenshot shows a Windows desktop environment with the RStudio application open. The window title is "RStudio". The interface includes several panes:

- Code Editor:** Displays two R script files: "intro.lecture.Rmd" and "process_twitter_translate.R". The code in "process_twitter_translate.R" is as follows:

```
38 #mfa_norway_china <- read_csv("mfa_norway_china.csv")
39
40 #Some word embeddings
41 path_to_data <- 'C:/Users/neilke/Dropbox/MFA_twitter/word_embeddings/'
42
43 # (glove) pre-trained embeddings
44 pre_trained <- readRDS(paste0(path_to_data, 'glove.rds'))
45
46 # transformation matrix
47 transform_matrix <- readRDS(paste0(path_to_data, 'khodakA.rds'))
48
49 # find contexts for regime type
50 contextD <- get_context(x = mfa_norway_china$translatedtext[mfa_norway_chi
51             window = 6, valuetype = "fixed", case_insensitive =
52             hard_cut = FALSE, verbose = FALSE)
53
54 contextA <- get_context(x = mfa_norway_china$translatedtext[mfa_norway_chi
55             window = 6, valuetype = "fixed", case_insensitive =
56             hard_cut = FALSE, verbose = FALSE)
```

- Console:** Shows the command "y <- "NUPI is the best!"
- Environment:** Shows the global environment with objects:

 - mfa_tweets_democ_ 531025 obs. of 49 variables
 - mfa_tweets_final 806494 obs. of 49 variables
 - pre_trained Large matrix (120000000 elements, 987 MB)
 - revolution 9 obs. of 2 variables
 - transform_matrix Large matrix (90000 elements, 720.2 kB)

- Files:** Shows the file structure under "C:\Users\neilke\Dropbox":

 - New Folder
 - Delete
 - Rename
 - More
 - ...
 - ...
 - Name
 - Size
 - Modified

Name	Size	Modified
RDData	2.9 KB	Nov 13, 2020, 11:10 PM
RHistory	10.9 KB	May 2, 2021, 10:52 AM
02.png	227 KB	Oct 19, 2020, 5:17 PM
04.png	222.4 KB	Oct 19, 2020, 5:17 PM
06.png	217.3 KB	Oct 19, 2020, 5:17 PM
08_and_17.png	258 KB	Oct 19, 2020, 5:18 PM

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

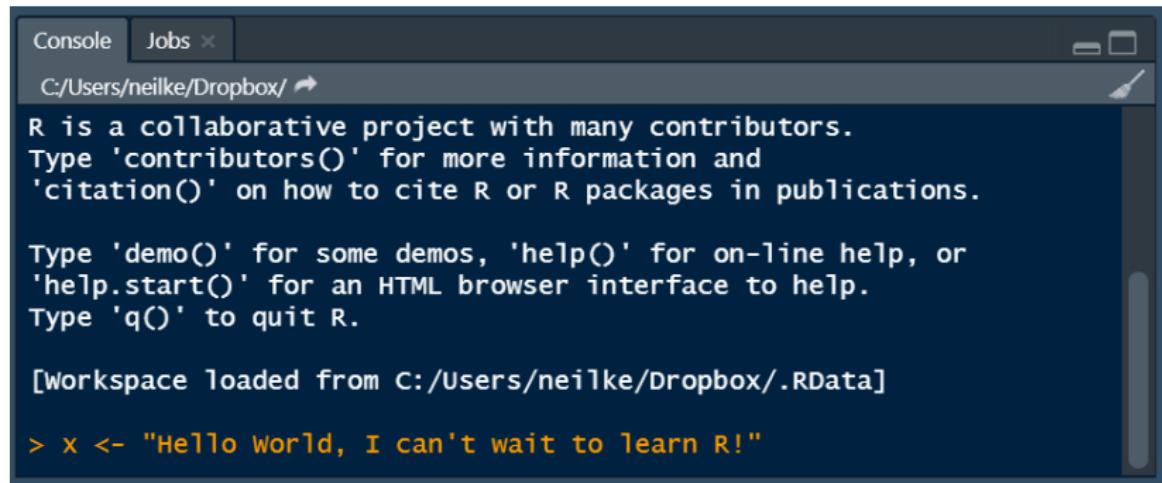
```
x <- "Hello World, I can't wait to learn R!"
```

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 . . .



The screenshot shows a dark-themed R console window. At the top, there are tabs for "Console" and "Jobs", with "Console" being active. Below the tabs, the current working directory is listed as "C:/Users/neilke/Dropbox/". The main area displays the standard R startup message, which includes information about contributors, citation methods, help resources, and how to quit. At the bottom of the message, it says "[Workspace loaded from C:/Users/neilke/Dropbox/.RData]". A single line of user code, "`> x <- "Hello world, I can't wait to learn R!"`", is visible at the bottom of the window.

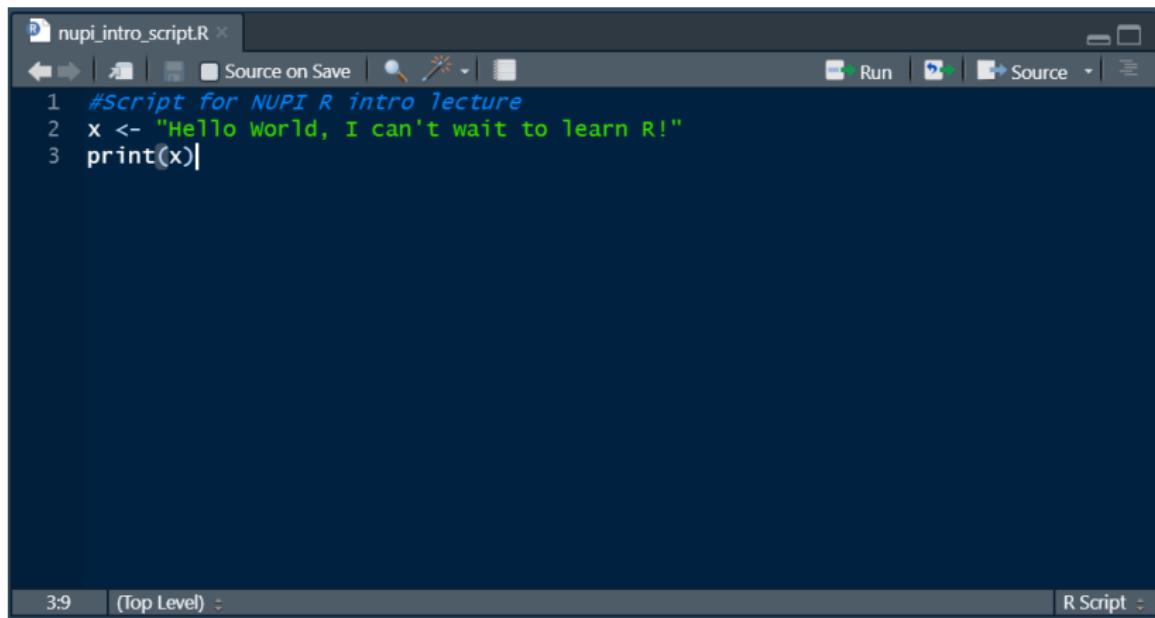
```
Console Jobs ×
C:/Users/neilke/Dropbox/ ↵
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from C:/Users/neilke/Dropbox/.RData]

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

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

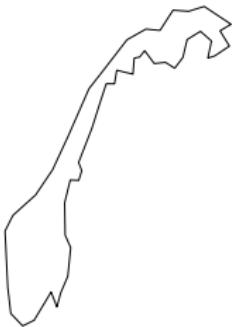


The screenshot shows the RStudio interface with an R script file open. The title bar reads "nupi_intro_script.R". The toolbar includes standard icons for back, forward, save, and run. The main editor area contains the following R code:

```
1  #Script for NUPI R intro lecture
2  x <- "Hello World, I can't wait to learn R!"
3  print(x)
```

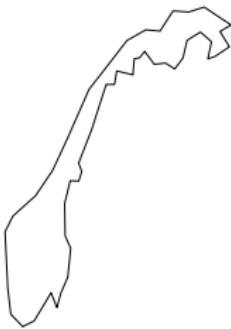
The status bar at the bottom shows "3:9" and "(Top Level)". The tab bar indicates the file is an "R Script".

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



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

Hello World, I can't wait to learn R!



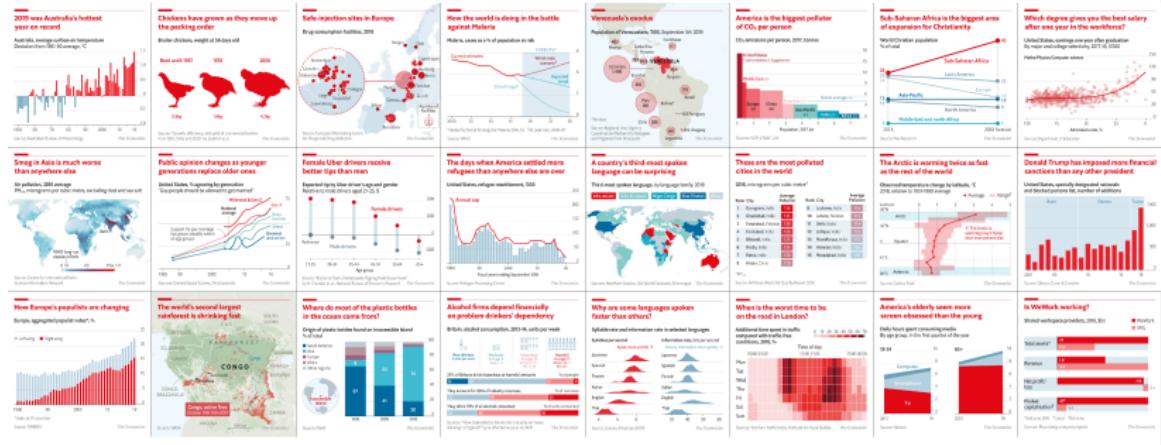
Beautiful graphics

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



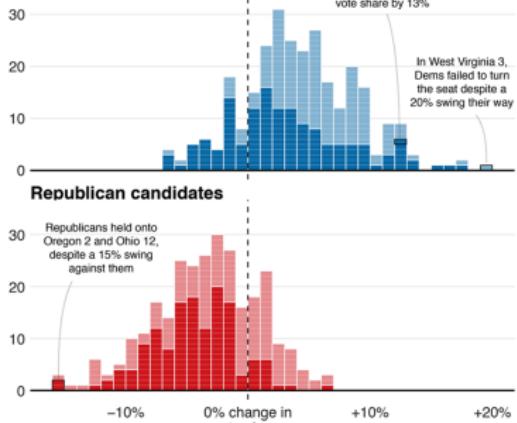
Recreate a graph from *The Economist* in ggplot2: http://rstudio-pubs-stat ic.s3.amazonaws.com/284329_c7e660636fec4a42a09eed968dc47f32.html

BBC R Cookbook

Blue wave

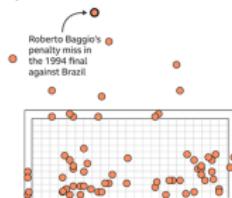
■ Won seat ■ Didn't win

Democrat candidates



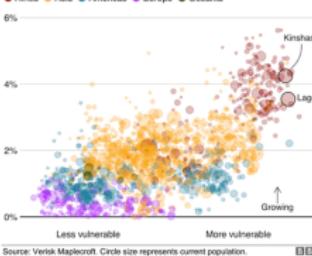
Where penalties are saved

World Cup shootout misses and saves, 1982-2014



Fast-growing cities face worse climate risks

Population growth 2018-2035 over climate change vulnerability

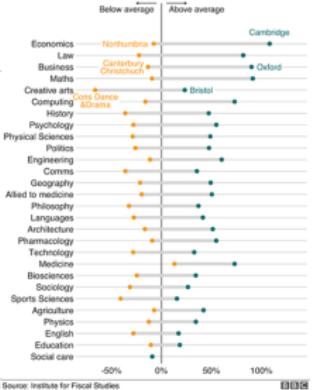


MPs rejected Theresa May's deal by 230 votes



Earnings vary across unis even within subjects

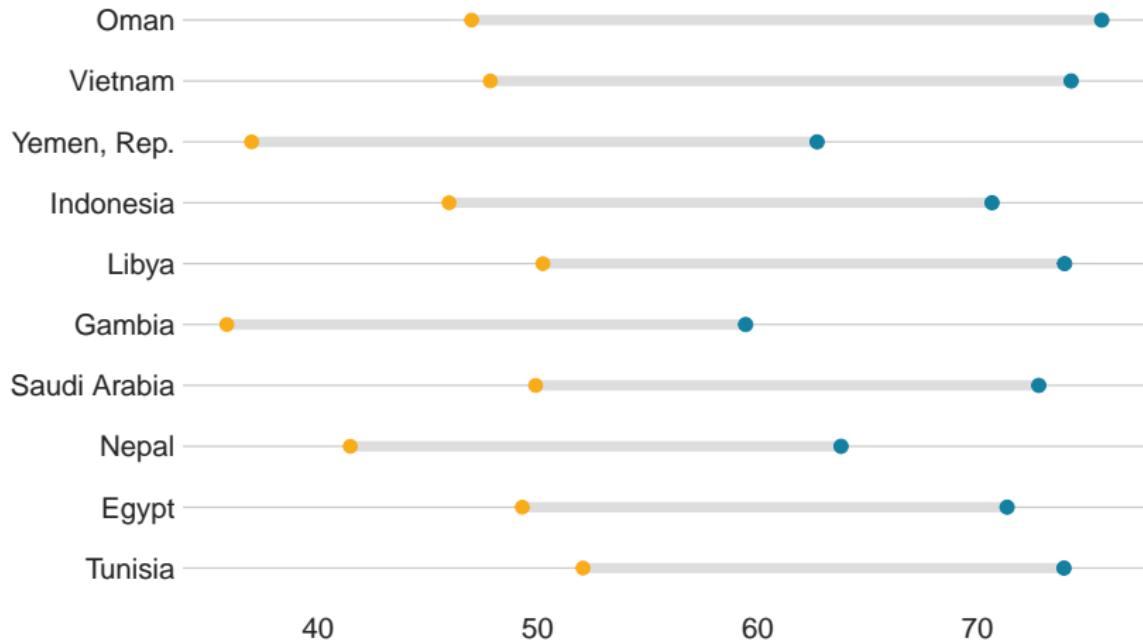
Impact on men's earnings relative to the average degree



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

We're living longer

Biggest life expectancy rise, 1967–2007



```
library("ggalt")
library("tidyverse")
#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")
```

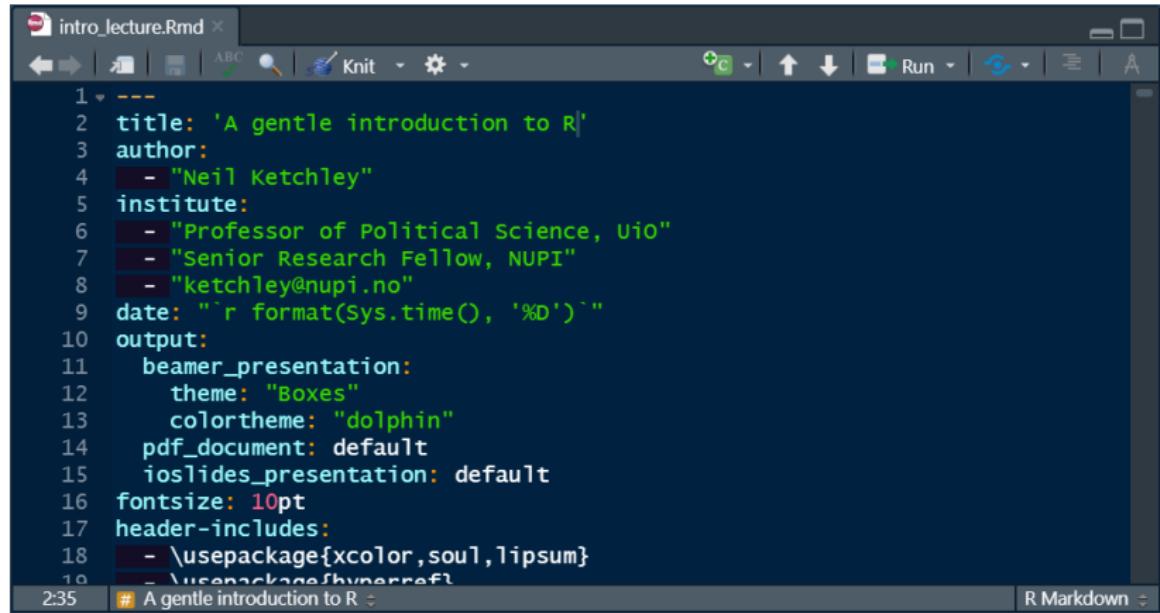
Disseminating reproducible research

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!



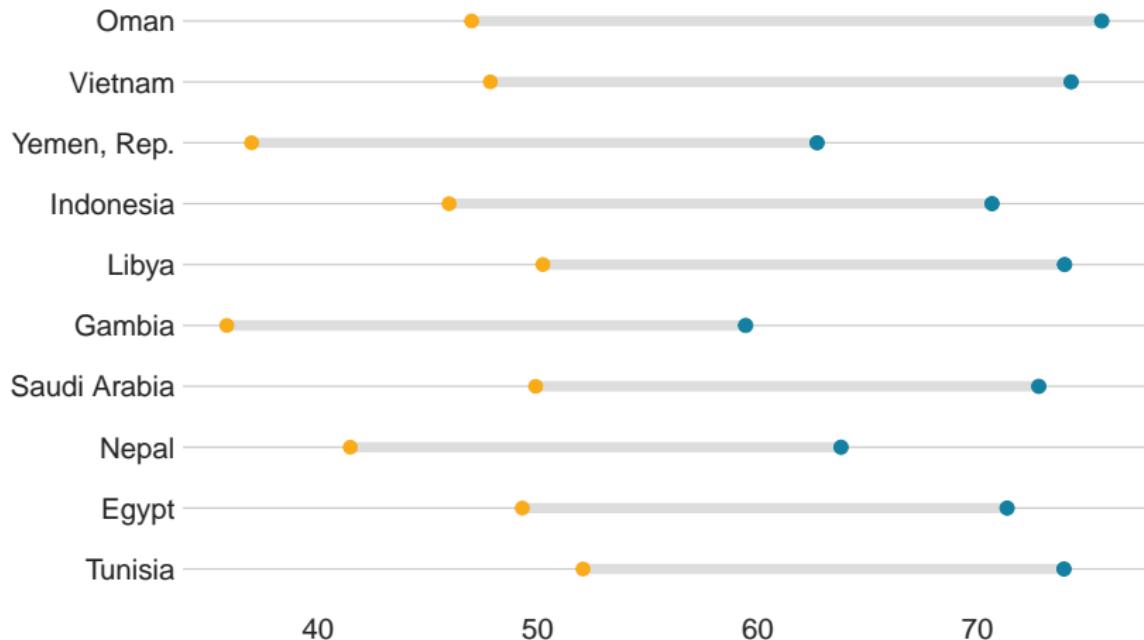
```
1 ---  
2 title: 'A gentle introduction to R'  
3 author:  
4 - "Neil Ketchley"  
5 institute:  
6 - "Professor of Political Science, UiO"  
7 - "Senior Research Fellow, NUPI"  
8 - "ketchley@nupi.no"  
9 date: "`r format(Sys.time(), '%D')`"  
10 output:  
11   beamer_presentation:  
12     theme: "Boxes"  
13     colortheme: "dolphin"  
14   pdf_document: default  
15   ioslides_presentation: default  
16 fontsize: 10pt  
17 header-includes:  
18 - \usepackage{xcolor,soul,lipsum}  
19 - \usepackage{hyperref}
```

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

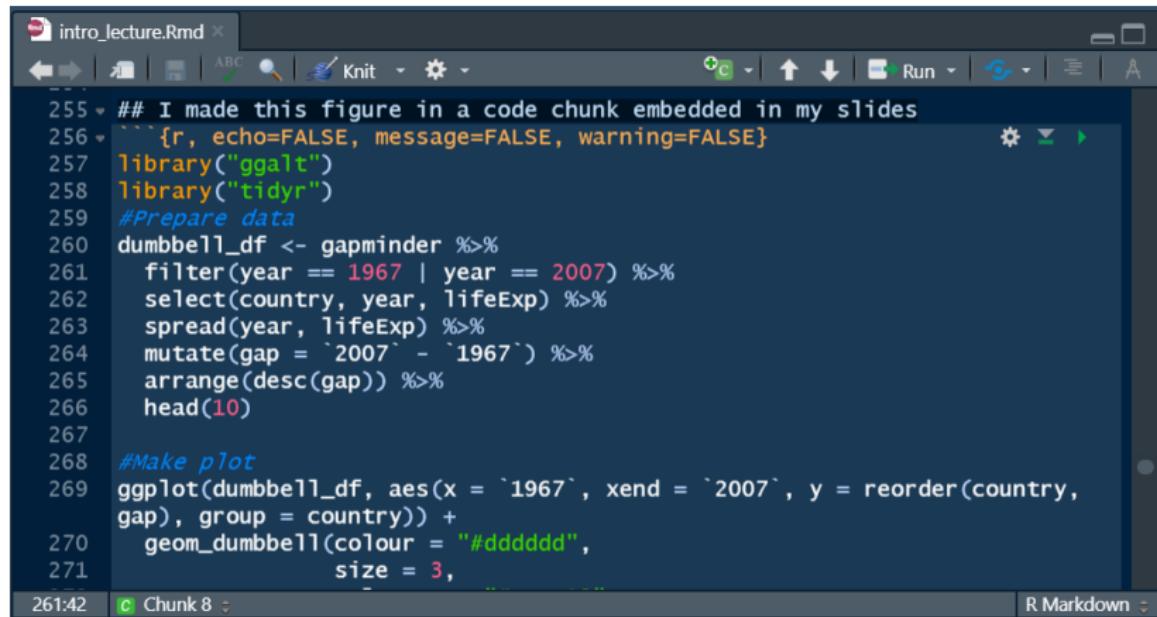
I made this figure in a code chunk embedded in my slides

We're living longer

Biggest life expectancy rise, 1967–2007



This means you can update your figures on the fly



The screenshot shows an RStudio interface with the following details:

- Title Bar:** intro_lecture.Rmd
- Toolbar:** Includes back, forward, file, ABC, search, Knit, settings, and other standard icons.
- Code Editor:** Displays R code for generating a dumbbell plot. The code includes library imports for ggalt and tidyverse, data preparation steps like filtering years 1967 and 2007, calculating a gap variable, arranging the data by country, and finally plotting it with geom_dumbbell. A note at the top indicates the figure was made in a code chunk.
- Status Bar:** Shows "261:42" (line number), "Chunk 8" (highlighted in green), and "R Markdown".

```
255 ## I made this figure in a code chunk embedded in my slides
256 ````{r, echo=FALSE, message=FALSE, warning=FALSE}
257 library("ggalt")
258 library("tidyverse")
259 #Prepare data
260 dumbbell_df <- gapminder %>%
261   filter(year == 1967 | year == 2007) %>%
262   select(country, year, lifeExp) %>%
263   spread(year, lifeExp) %>%
264   mutate(gap = `2007` - `1967`) %>%
265   arrange(desc(gap)) %>%
266   head(10)
267
268 #Make plot
269 ggplot(dumbbell_df, aes(x = `1967`, xend = `2007`, y = reorder(country,
270   gap), group = country)) +
271   geom_dumbbell(colour = "#dddddd",
272                 size = 3,
```

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!

The screenshot shows the RStudio interface with two files open in the left panel: "intro_lecture.Rmd" and "blinded_manuscript.RMd". The "blinded_manuscript.RMd" file is the active tab. The code editor contains the following R Markdown code:

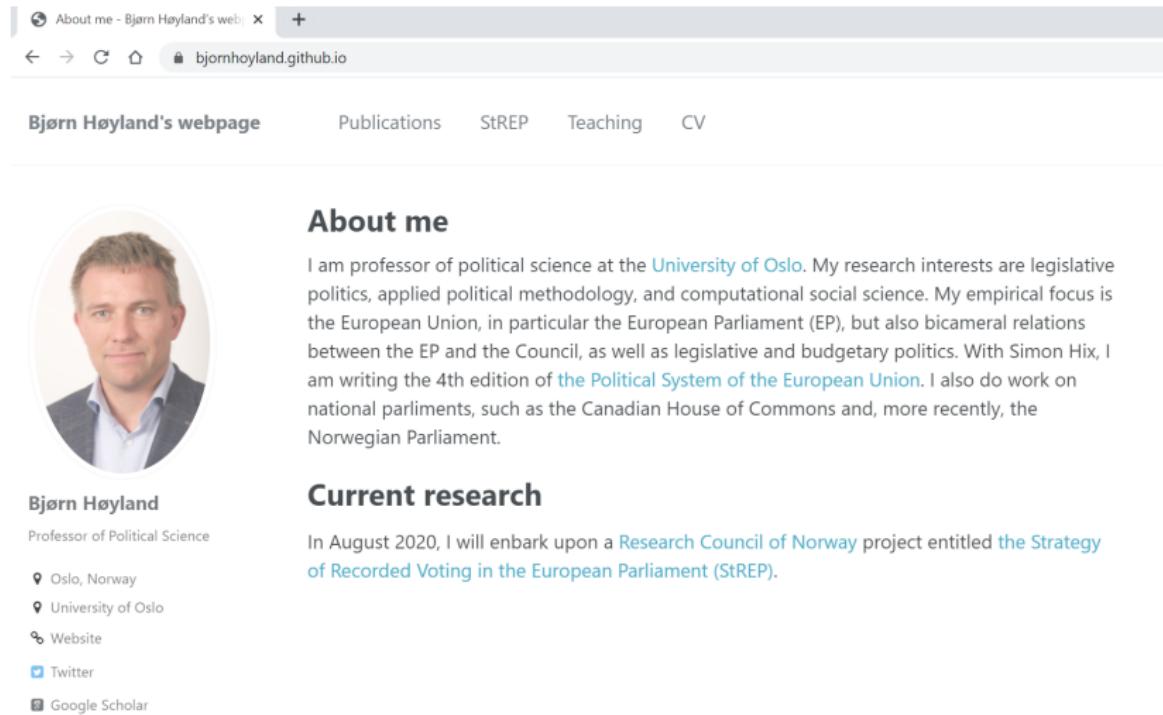
```
47
48 \newpage
49 \clearpage
50 \pagenumbering{arabic}
51
52 # Introduction
53
54 How should we measure terrorism? The answer seems obvious: count the
  attacks. But attacks are only the subset of terrorist activity that
  happen to reach execution. Many plots are foiled by police or derail for
  other reasons, never to enter terrorism datasets. In this article, we
  build on recent advances in plot data collection to conduct the first
  systematic examination of this potential measurement problem. We show
  that, while plots and attacks are often highly correlated, they differ
  enough to produce statistically different results in common study
  designs. Answers to the question "what causes terrorism" can therefore
  vary depending on how we operationalize the notion of terrorist activity.
55
56 I The inquiry matters because attack counts are central to the
  # Plots, Attacks, and the Measurement of terrorism
```

The preview pane on the right shows the rendered content of the manuscript.

Check out the article options in R Markdown:

<https://bookdown.org/yihui/rmarkdown/journals.html>

My colleagues make their websites in R Blogdown!



A screenshot of a website for "Bjørn Høyland's webpage". The header includes a navigation bar with links to "About me", "Publications", "StREP", "Teaching", and "CV". Below the header is a circular profile picture of a man (Bjørn Høyland) and his name. The main content area has two sections: "About me" and "Current research".

About me

I am professor of political science at the [University of Oslo](#). My research interests are legislative politics, applied political methodology, and computational social science. My empirical focus is the European Union, in particular the European Parliament (EP), but also bicameral relations between the EP and the Council, as well as legislative and budgetary politics. With Simon Hix, I am writing the 4th edition of [the Political System of the European Union](#). I also do work on national parliments, such as the Canadian House of Commons and, more recently, the Norwegian Parliament.

Current research

In August 2020, I will embark upon a [Research Council of Norway](#) project entitled [the Strategy of Recorded Voting in the European Parliament \(StREP\)](#).

Bjørn Høyland

Professor of Political Science

📍 Oslo, Norway
📍 University of Oslo
🔗 Website
🔗 Twitter
🔗 Google Scholar

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

Books about R are made in R Markdown . . .

R Markdown: The Definitive Guide

Routledge Taylor & Francis Group

CRC Press Taylor & Francis Group

Subjects Our Customers Our Products Blog SALE Products Search by keywords, subject, or ISBN

30% OFF EBOOKS

1st Edition

R Markdown
The Definitive Guide

By Yihui Xie, J.J. Allaire, Garrett Grolemund

Copyright Year 2019

Paperback £28.99 Hardback £69.99 eBook £20.29

ISBN 9781138359338
Published July 18, 2018 by Chapman and Hall/CRC
338 Pages

Format Paperback Quantity 1 GBP £28.99

Enlarge Download

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

The screenshot shows a web browser window with the title "R Markdown: The Definitive Guide" at the top. The URL in the address bar is "bookdown.org/yihui/rmarkdown/". The page content includes a sidebar on the left with a table of contents for the "Preface" chapter, listing sections like "How to read this book", "Structure of the book", "Software information and conventions", "Acknowledgments", "About the Authors", "Yihui Xie", "J.J. Allaire", "Garrett Grolemund", "Get Started", "Installation", "Basics", "Example applications", "Airbnb's knowledge repo...", "Homework assignments ...", "Personalized mail", and "2017 Employer Health B...". The main content area features the title "R Markdown: The Definitive Guide" in large bold letters, followed by author credits ("Yihui Xie, J. J. Allaire, Garrett Grolemund") and a date ("2021-04-09"). Below the title, there is a section titled "Preface" with a note about the book's publication by Chapman & Hall/CRC and its Creative Commons license. At the bottom right, there is a logo for "The R Series" with "R Markdown" below it.

R Markdown: The Definitive Guide

Yihui Xie, J. J. Allaire, Garrett Grolemund

2021-04-09

Preface

Note: This book has been published by Chapman & Hall/CRC. The online version of this book is free to read here (thanks to Chapman & Hall/CRC), and licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The R Series

R Markdown

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")

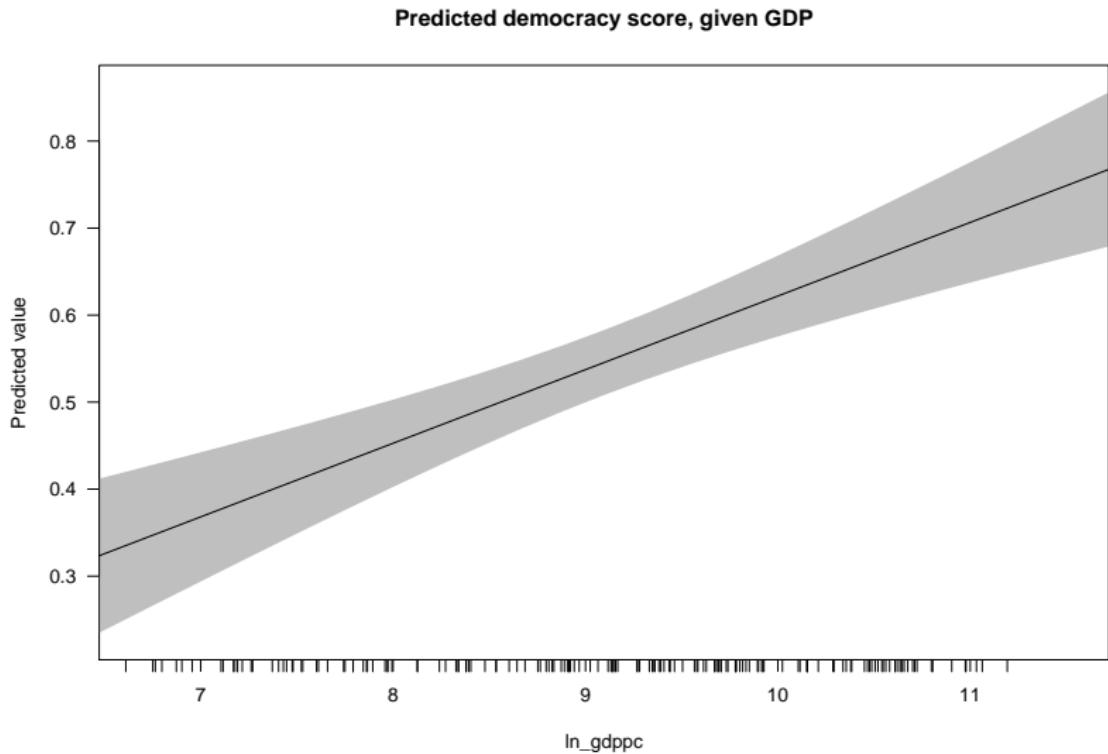
```

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.

Federal Register :: Office of the C X +

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

Office of the Chief of Protocol; Gifts to Federal Employees From Foreign Government Sources Reported To Employing Agencies in Calendar Year 2017

A Notice by the State Department on 03/07/2019

PUBLISHED DOCUMENT

The Office of the Chief of Protocol, Department of State, submits the following comprehensive listing of the statements which, as required by law, federal employees filed with their employing agencies during calendar year 2017 concerning gifts received from foreign government sources. The compilation includes reports of both tangible gifts and gifts of travel or travel expenses of more than minimal value, as defined by the statute. Also included are gifts received in previous years including one gift in 2011, four gifts in 2012, two gifts in 2013, two gifts in 2014, seven gifts in 2015, and one gift with an unknown date. These latter gifts are being reported in this year's report as the Office of the Chief of Protocol, Department of State, did not receive the relevant information to include them in earlier reports. Any agency not listed in this report either did not

DOCUMENT DETAILS

Printed version: PDF

Publication Date: 03/07/2019

Agency: Department of State

Document Type: Notice

Document Citation: 84 FR 8361

Page:

Feedback

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

The screenshot shows a web browser displaying a Federal Register document from the Office of the Chief of Protocol. The URL is federalregister.gov/documents/2019/03/07/2019-04063/office-of-the-chief-of-protocol-gifts-to-federal-employees-from-foreign-government-.... The page title is "President".

PUBLISHED DOCUMENT			
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	Sayid 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

Federal Register :: Office of the C X +

federalregister.gov/documents/2019/03/07/2019-04063/office-of-the-chief-of-protocol-gifts-to-federal-employees-from-foreign-government... ☆ 📁 🌐 🛡️ 🎯 🚧 🌐 🌐 🌐

Agency: The White House—Executive Office of the President
[Report of Tangible Gifts Furnished by the White House—Executive Office of the President]

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	Statue, carved Ohio sandstone, in 2 pieces, depicting standing male lion wearing crown, over cross and orb, rectangular base. Rec'd 2/13/17. Est. Value—\$450.00. Disposition— View Details	The Right Honorable Justin Trudeau, P.C., M.P., Prime Minister of Canada	Non-diplomatic would embarrass U.S. Government

table 584 x 22546
Margin 0px 0px 20px
Padding 10px
ACCESSIBILITY
Name Agency: The White House—Executive O...
Role Rec'd Pedro Pablo Kuczynski
Keyboard-focusable 50.00.

Feedback

Elements Console Sources Network Performance

```
<!DOCTYPE html>
<html class="mdrnzr-js mdrnzr-cssgradients mdrnzr-svg"> == $o
> <head></head>
<body id="documents" class="show has_js" data-environment="production" data-honeybadger-js-api-key="229f39c0" style="overflow: hidden auto;">
  <script nonce="has_js">
    document.getElementsByTagName('body')[0].className += ' has_js'
  </script>
  <a href="#" title="Skip to Content" class="skip_to_content">
    Skip to Content
  </a>
  <div class="header" id="header_refresh">...</div>
  <div id="main">
    <div id="printDisclaimer">...</div>
  ... html.mdnrzr.js.mdnrzr-cssgradients.mdnrzr-svg

```

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

:how .cls +

element.style {

```
html, body, div, span, object, iframe, h1, h2, h3, h4, application_0258d.css:1
h5, h6, p, blockquote, pre, abbr, address, cite, code, del, dfn, em, img, ins, kbd, q, samp, small, strong, sub, sup, var, b, i, dl, dt, dd, ol, ul, li, fieldset, form, label, legend, table, caption, tbody, tfoot, thead, tr, th, td, article, aside, canvas, details, figcaption, figure, footer, header, hgroup, menu, nav, section, summary, time, mark, audio, video {
  margin: 0px;
  padding: 0px;
  border: 0px;
  outline: 0px;
  font-size: 100%;
  vertical-align: baseline;
  background: transparent;
}
```

user agent: chrome/69.0.3497.100

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!

intro_lecture.Rmd process_US_gifts.R gifts_2017a

Filter

	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	Identifier
18	The Honorable Donald J. Trump, President of the United Sta...	Plaque, American flag, in natural color stone, 16'h x 24"w, a...	H
19	The Honorable Donald J. Trump, President of the United Sta...	Engraving, black ink, titled "Misericordiae vltvs", 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. 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 via 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 **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, 1010/— Between Cairo, Alexandria, **and** Port Said various stations burnt signal-boxes **destroyed**, and rails and sleepers removed **wong** the main and subsidiary lines. One train arrived at Cairo from 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 **destroyed** in all directions. There is no communication other than by wireless telegraph and aeroplane **between** 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_telegram.json")
```

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

OCR with daiR

1 Sir M. Cheetham to Earl Curzon.—(Received March 19.)
134 15 16 118
1920 No. 406. Urgent! 222
Telegraphic.
FOLLOWING is summary of events reported during the twenty-four hours ending noon, 18th March, 1919. 23 2425 26 2728 29
30 31 32 33 34 35 36 37 38 390 41
42 43 4445 46 47 48
Cairo, March 18, 1919.
49 50 51 52 53 54 55 56 57 58 59 60 662 63
64 Between Cairo, Alexandria and Port Said various stations burnt signal-boxes 6566 67 68 69 70 71 72 73 74 75 76 77
destroyed and rails and sleepers removed along the main and subsidiary lines.
78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93
One train arrived at Cairo after journey of twenty-six hours, preceded by military train repairing damage to communications with Upper Egypt.
94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 1090 111 112 113 114 115 116 117 118 119 120
beyond Wasta where serious damage to other stations destroyed. Similar damage further south reported by telegram. Sudan natives constantly engaged in 121 122 123 124 125 126 127 128 129 130 131 132 133
destroying lines in numerous districts until disrupted fire.
144 145 146 147 148 149 150 151 152 153
154 There is no communication other than by wireless telegraph and aeroplane 155 156 157 158 159 160 161 162 163 164
165 between Cairo and the provinces.
171 172 173 174 175 176 177 178 179 180 181 182 183 184 185
186 At Cairo a peaceful demonstration in which some thousands took part headed by university intellectuals took place yesterday without incident except where once 187 188 189 190 191 192 193 194 195 196 197
198 interfered with by party soldiers and left with casualties.
210 Demonstration this morning at Ismailia was dispersed with several casualties. 211 212 213 214 215 216 217 218 219 220
221 222 Further demonstration in Cairo was ordered to disperse by military authorities 223 224 225 226 227 228 229 230 231
232 233 234 235
236 237 238 239 240 241 242 243 244 245 246 247
248 249 250 251 252 253 254 255 256 257 258 259
260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275
276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291
various sugar plantations were attacked but precautions taken. Attacks were made on sugar factories at Cairo, but assailants beaten off.

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 via 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

The screenshot shows a web browser window displaying the "Welcome" page of the "R for Data Science" website. The URL in the address bar is r4ds.had.co.nz. The page content includes a search bar, a table of contents, and a main welcome text. To the right, there is a sidebar with links like "On this page", "Welcome", "Acknowledgements", "View source", and "Edit this page". A book cover for "R for Data Science" by Hadley Wickham and Garrett Grolemund is displayed, featuring a green parrot.

R for Data Science

Search

Table of contents

Welcome

1 Introduction

Explore

2 Introduction

3 Data visualisation

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

Wrangle

9 Introduction

10 Tibbles

11 Data import

12 Tidy data

On this page

Welcome

Acknowledgements

View source

Edit this page

Welcome

This is the website for "**R for Data Science**". This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

This website is (and will always be) **free to use**, and is licensed under the Creative

OREILLY

R for Data Science

VISUALIZE, MODEL, TRANSFORM, Tidy, AND IMPORT DATA

Hadley Wickham & Garrett Grolemund

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

ECPR Summer and Winter Schools

The screenshot shows a web browser window with the URL ecpr.eu/SummerSchool. The page title is "Virtual Summer School". Below the title, there is a sub-headline "Tackle your research with confidence", the date "2 – 20 August 2021", and a Twitter handle "#ecprvms21". A call-to-action button says "Funding applications now open! Deadline Sunday 6 June.". To the right, a sidebar menu lists "Funding Opportunities", "How it works", "Course fees", and "Guides" (which is highlighted in black). Other menu items include "Code of Conduct" and "The Loop: ECPR's political science blog".

We're delighted to announce that our next Summer School will be held virtually, from 2 – 20 August 2021.

Offering a comprehensive virtual programme of courses on variance-based, case-based and interpretive methods and techniques, our Virtual Methods School allows you to select and combine courses to meet the demands of your project. The result? A rounded training experience giving you the tools you need to tackle your research with confidence.

How it works

- A three-week programme
- Expertly selected courses perfectly tailored to suit the digital environment
- 10+ hours of live class per course, plus pre-recorded content, readings, social spaces and more

What to expect

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

Try out the exercises yourself



GitHub

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/>