

Part 1

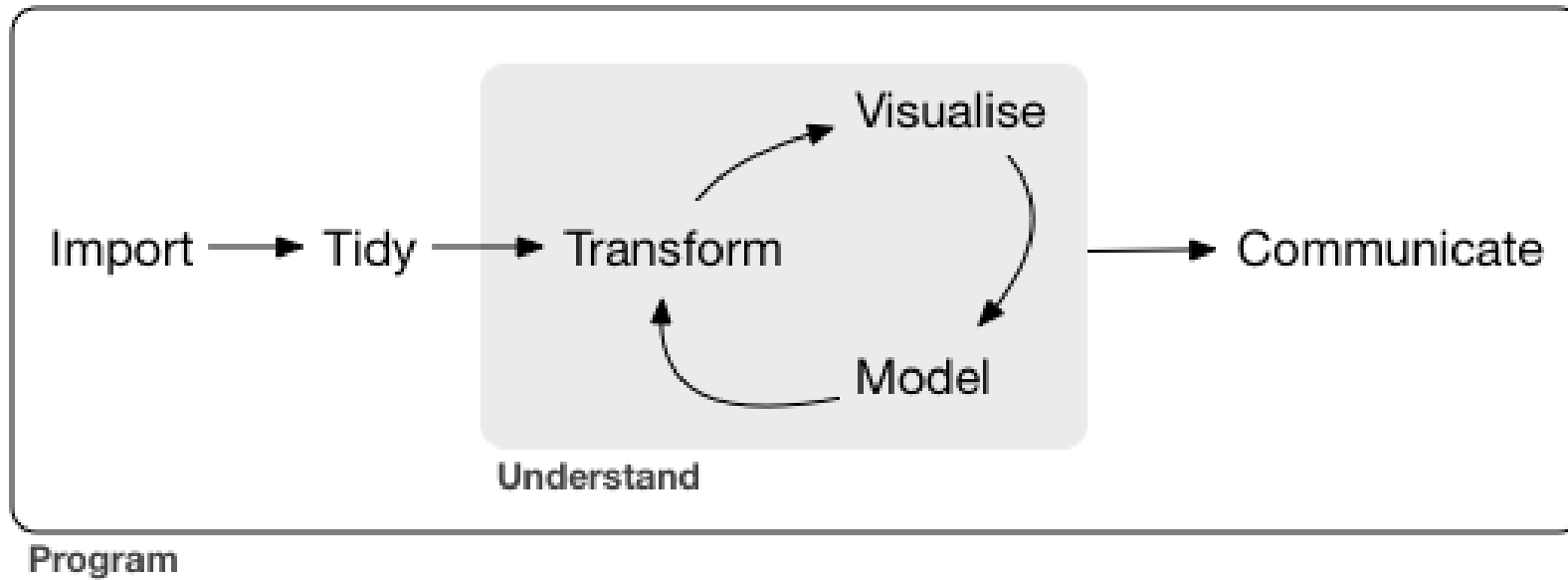


```
Session Title.R x
← → | ↺ | 💾  Source on Save | 🔍 ✨ | 📄 ↶ ↷ | ➡ Source
1 Session <- "Dive Headfirst into R"
2 With <- c("Dr. Jun Zhang",
3         "Dr. Ken Blake")
4 Date <- as.Date("03/02/2023",
5               "%m/%d/%y")
6 print(Session)
7 print(With)
8 print(Date)
9
10:1 (Top Level) R Script
```



You don't have to remember every single R command.

Feel free to Google it when you need to!



Data science with R workflow

R & RStudio

The R programming language is widely used among statisticians and data miners for statistical computing and graphics.

R is a software environment to process R programming languages.

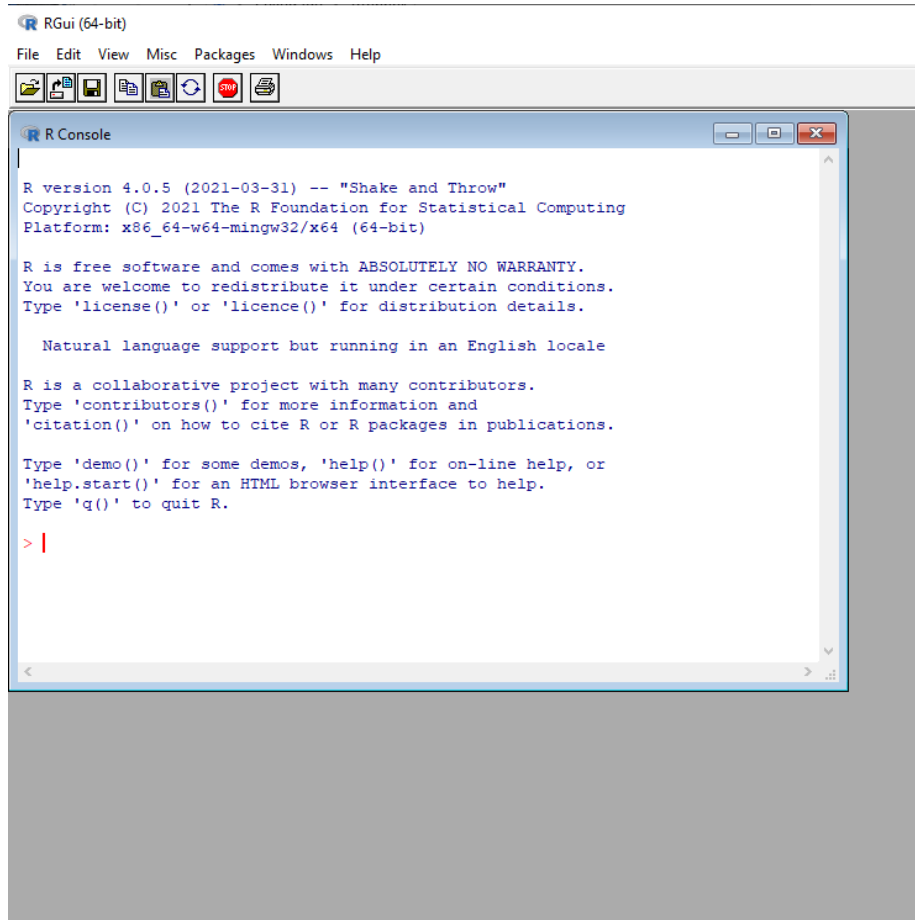
RStudio is an integrated development environment for R programming.

R & RStudio

If R is the engine and bare bones of your car, then RStudio is like *the rest of the car*. The engine is super critical part of your car. But in order to make things properly functional, you need to have a steering wheel, comfy seats, a radio, rear and side view mirrors, storage, and seatbelts.

RMarkdown for Scientists *Nicholas Tierney*

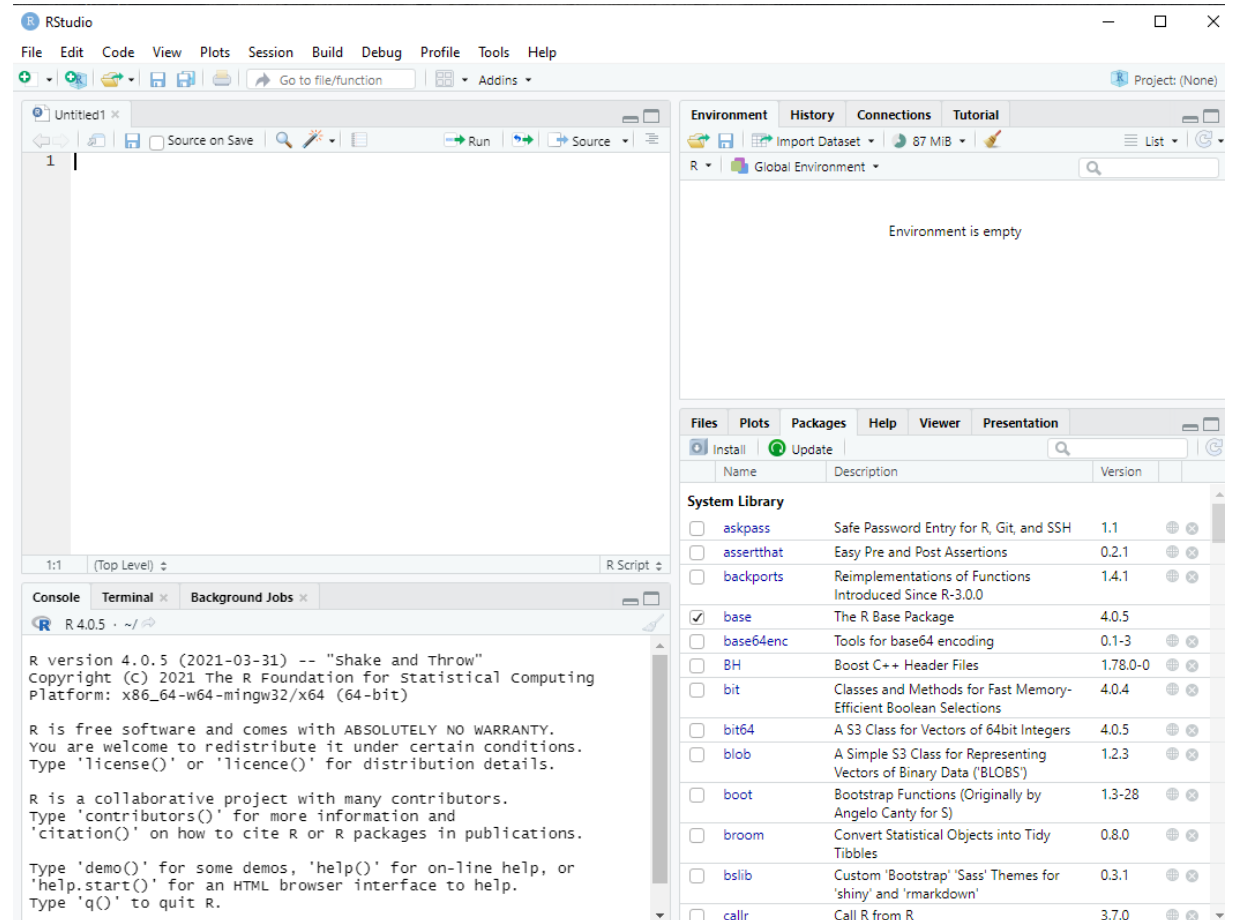
R & RStudio



RGui (64-bit)
File Edit View Misc Packages Windows Help

```
R Console  
R version 4.0.5 (2021-03-31) -- "Shake and Throw"  
Copyright (C) 2021 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
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.  
  
> |
```

R interface



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

```
Untitled1 x  
1 |  
  
Console Terminal x Background Jobs x  
R 4.0.5 · ~/ |  
R version 4.0.5 (2021-03-31) -- "Shake and Throw"  
Copyright (C) 2021 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
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.
```

Environment History Connections Tutorial
R Global Environment
Environment is empty

Name	Description	Version
<input type="checkbox"/> askpass	Safe Password Entry for R, Git, and SSH	1.1
<input type="checkbox"/> assertthat	Easy Pre and Post Assertions	0.2.1
<input type="checkbox"/> backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
<input checked="" type="checkbox"/> base	The R Base Package	4.0.5
<input type="checkbox"/> base64enc	Tools for base64 encoding	0.1-3
<input type="checkbox"/> BH	Boost C++ Header Files	1.78.0-0
<input type="checkbox"/> bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
<input type="checkbox"/> bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
<input type="checkbox"/> blob	A Simple S3 Class for Representing Vectors of Binary Data (BLOBS)	1.2.3
<input type="checkbox"/> boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
<input type="checkbox"/> broom	Convert Statistical Objects into Tidy Tibbles	0.8.0
<input type="checkbox"/> bslib	Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'	0.3.1
<input type="checkbox"/> callr	Call R from R	3.7.0

RStudio interface

Packages

An R package is a collection of functions, data, and documentation that extends the capabilities of base R.

The packages installed are not loaded by default.

You will not be able to use the functions, objects, and help files in a package until you load it.



R

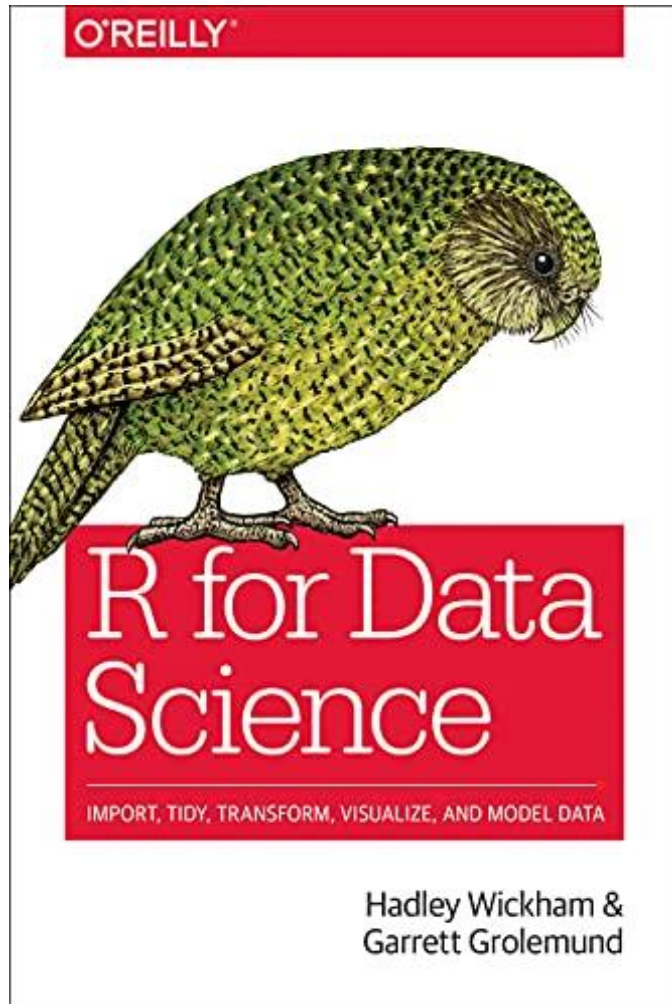


R packages

Packages

- **tidyverse:** include all the packages required in the data science workflow, ranging from data exploration to data visualization.
 - Data Visualization and Exploration: ggplot2
 - Data Wrangling and Transformation: dplyr, tidyr, stringr, forcats
 - Data Import and Management: tibble, readr
 - Functional Programming: purr
- **lubridate:** helps users to easily manipulate date and time data. It provides tools for parsing, formatting and manipulating dates and times.
- **rtweet:** collect and organize Twitter data via Twitter's REST and stream API
- **plotly:** creating interactive web-based graphs via the open source JavaScript graphing library plotly.js

Resources



<https://posit.cloud/>

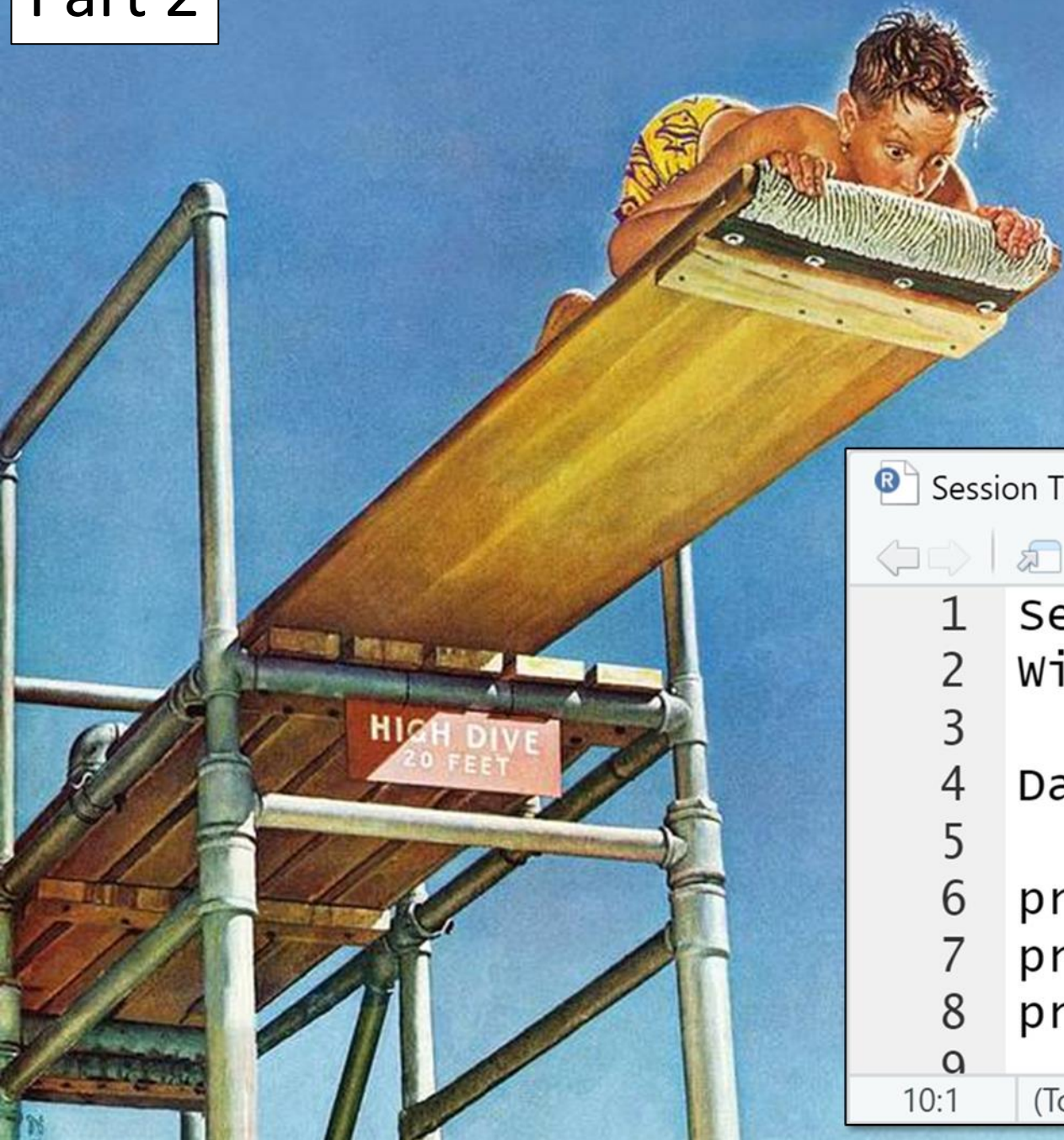
Learning statistics with R -
<https://learningstatisticswithr.com/book/>

R for Data Science - <https://r4ds.had.co.nz/>

Data Visualization - <https://socviz.co/>

<https://www.geeksforgeeks.org/>

Part 2



```
Session Title.R x
Source on Save
1 Session <- "Dive Headfirst into R"
2 With <- c("Dr. Jun Zhang",
3           "Dr. Ken Blake")
4 Date <- as.Date("03/02/2023",
5                "%m/%d/%y")
6 print(Session)
7 print(With)
8 print(Date)
9
10:1 (Top Level) R Script
```

```
#####  
if (!require("tidyverse")) install.packages("tidyverse")  
if (!require("readr")) install.packages("readr")  
if (!require("dplyr")) install.packages("dplyr")  
if (!require("tidytext")) install.packages("tidytext")  
library(tidyverse)  
library(readr)  
library(dplyr)  
library(tidytext)  
library(stringr) # Part of the tidyverse package
```

```
#####  
query <- "'Joe Biden'" #Enter search term(s)  
startdate <- "20230102" #Enter preferred start date  
enddate <- "20230219" #Enter preferred end date  
sources <- c("washingtonpost.com",  
            "nytimes.com") #Enter sources to search
```

```
#####  
#Generating a sequence of dates  
startdate2 <- as.Date(startdate, "%Y%m%d")  
enddate2 <- as.Date(enddate, "%Y%m%d")  
dates <- seq(as.Date(startdate2), as.Date(enddate2), "days")  
dates <- format(dates, "%Y%m%d")  
#Estimating run time for query  
Minutes <- round((length(sources)*(length(dates)*2.5/60)), digits = 1)  
Hours <- round((length(sources)*(length(dates)*2.5/3600)), digits = 1)
```

Plunge incentive:

Look at what this one script can do for you.

Specify search terms, dates, and sources ...

URL	MobileURL	Date	Title
1 https://www.washingtonpost.com/politics/new-documents-...	NA	2023-01-03 22:15:00	New documents detail Sen . Ron John
2 https://www.washingtonpost.com/national-security/2023/0...	NA	2023-01-03 21:00:00	The most intriguing revelations , new
3 https://www.washingtonpost.com/opinions/2023/01/03/mili...	NA	2023-01-03 20:30:00	Opinion Troops who refused covid va
4 https://www.washingtonpost.com/opinions/2023/01/03/kev...	NA	2023-01-03 20:15:00	Opinion Kevin McCarthy disastrous l
5 https://www.washingtonpost.com/politics/2023/01/03/miss...	NA	2023-01-03 17:30:00	Missed some headlines over the holid
6 https://www.washingtonpost.com/opinions/2023/01/03/ge...	NA	2023-01-03 15:15:00	Opinion What George Santos has in
7 https://www.washingtonpost.com/business/time-is-running...	NA	2023-01-03 14:30:00	Time Is Running Out for Afghan Refug
8 https://www.washingtonpost.com/politics/2023/01/03/bigg...	NA	2023-01-03 14:00:00	The biggest health stories to watch in
9 https://www.washingtonpost.com/business/trumps-tax-retu...	NA	2023-01-03 12:30:00	Trump Tax Returns Are Only Part of th
10 https://www.washingtonpost.com/business/big-tech-is-in-cr...	NA	2023-01-03 07:30:00	Big Tech Is in Crisis . That Exactly Wha
11 https://www.washingtonpost.com/politics/trump-lawyers-q...	NA	2023-01-04 22:45:00	Trump lawyers questioned Nevada 20.
12 https://www.washingtonpost.com/opinions/2023/01/04/mik...	NA	2023-01-04 20:30:00	Opinion Mikheil Saakashvili should n

Environment History Connections Tutorial

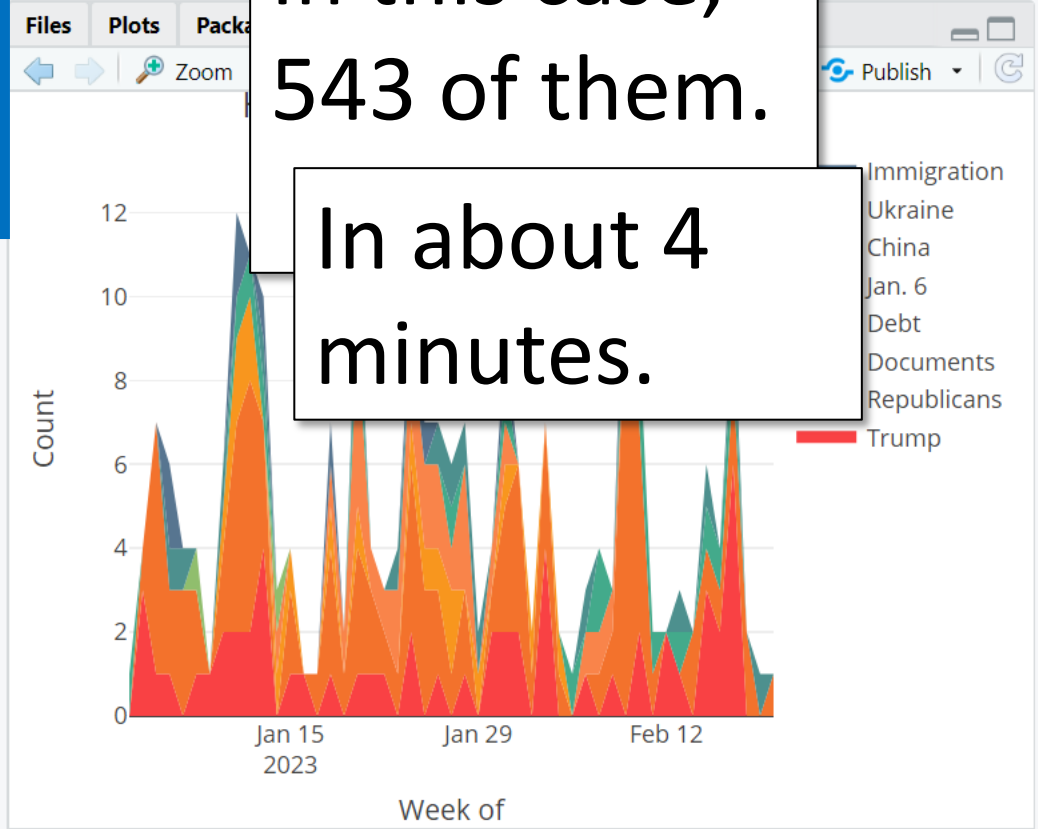
R | Global Environment

Data

- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig list of 8
- Headlines **543 obs. of 15 variables**
- WordCounts 1720 obs. of 2 variables

In this case, 543 of them.

In about 4 minutes.



... and get the URL, pub date & headline for every matching article.

```
showgrid = TRUE))
```

```
fig
```

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

Trump	Republicans	Documents	Debt	Jan6	China	Ukraine	Immigration	Day	WeekOf
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
1	1	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
1	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-03	2023-01-02
1	0	0	0	0	0	0	0	2023-01-05	2023-01-02
0	0	0	0	0	0	0	0	2023-01-05	2023-01-02

Showing 1 to 12 of 543 entries, 15 total columns

Environment History Connections Tutorial

Import Dataset 422 MiB

R Global Environment

Data

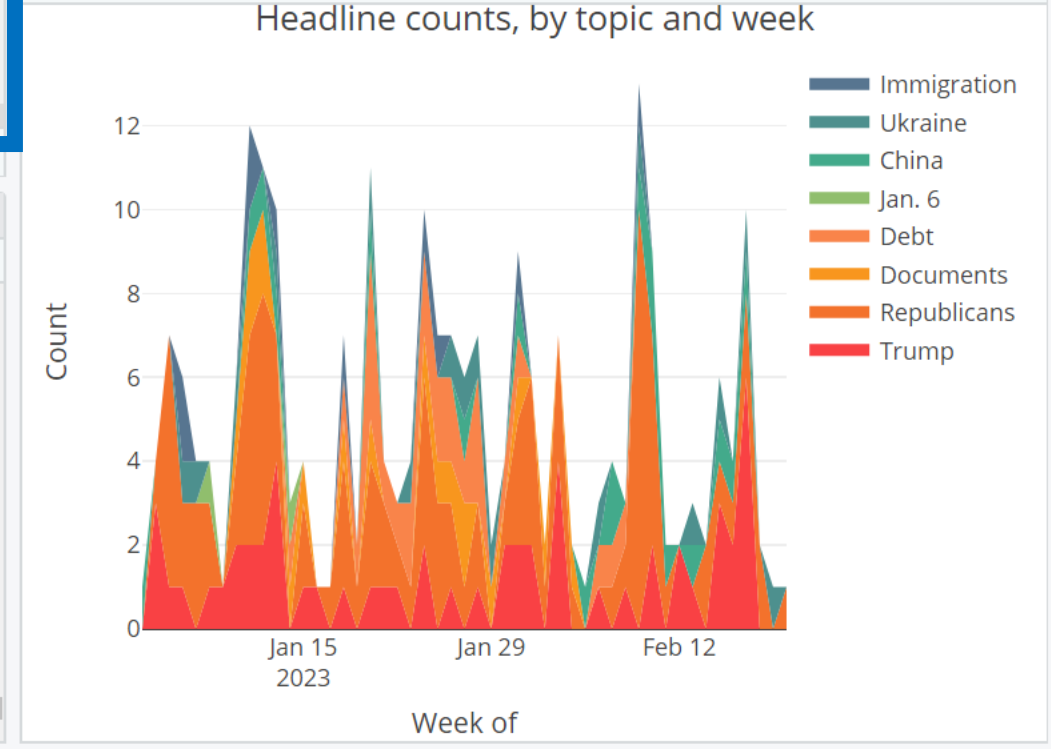
- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- wordCounts 1726 obs. of 2 variables

Values

Hours 0.1

Files Plots Packages Help Viewer Presentation

Zoom Export Publish



Headlines autocoded for mentions of user-defined topic keywords.

showgrid = TRUE))

fig

Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Showing 1 to 13 of 1,726 entries, 2 total columns

Environment History Connections Tutorial

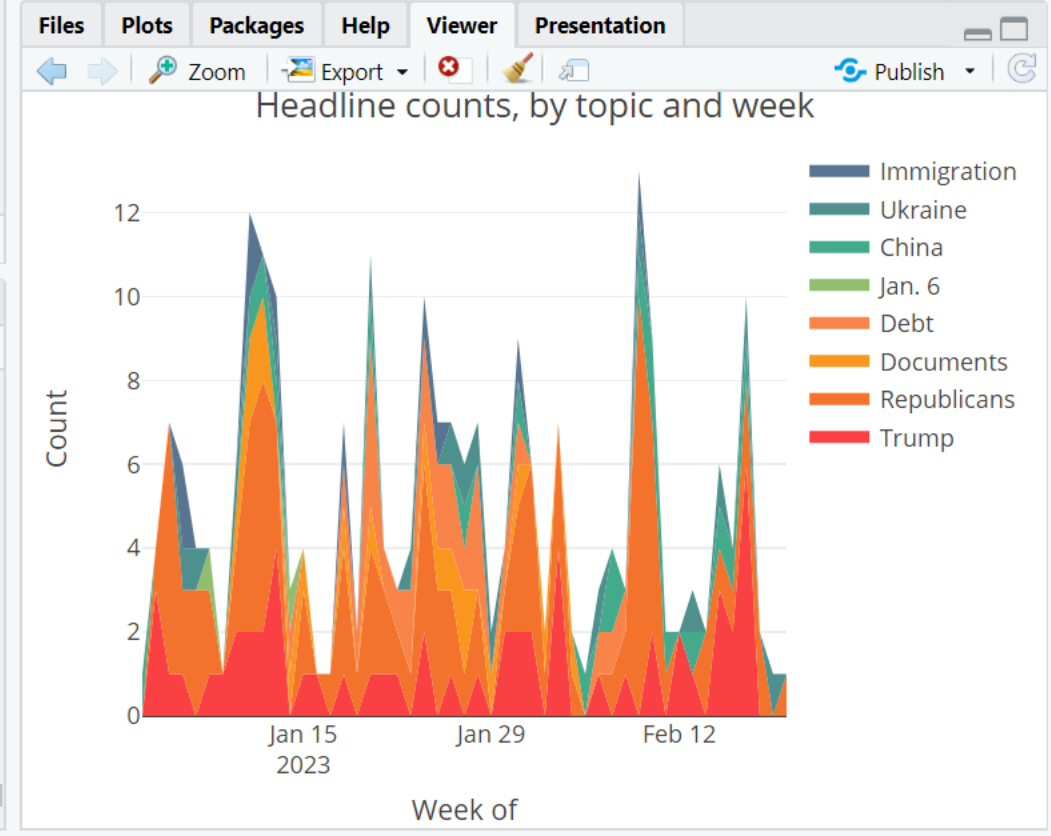
Import Dataset 422 MiB

R Global Environment

- AggByweek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- WordCounts 1726 obs. of 2 variables

Values

Hours	0.1
Minutes	4.1
searchterms	"border migra"



A headline word frequency analysis to help identify topics.

```
showgrid = TRUE))
```

```
fig
```

Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Showing 1 to 13 of 1,726 entries, 2 total columns

Environment History Connections Tutorial

Import Dataset 422 MiB

R Global Environment

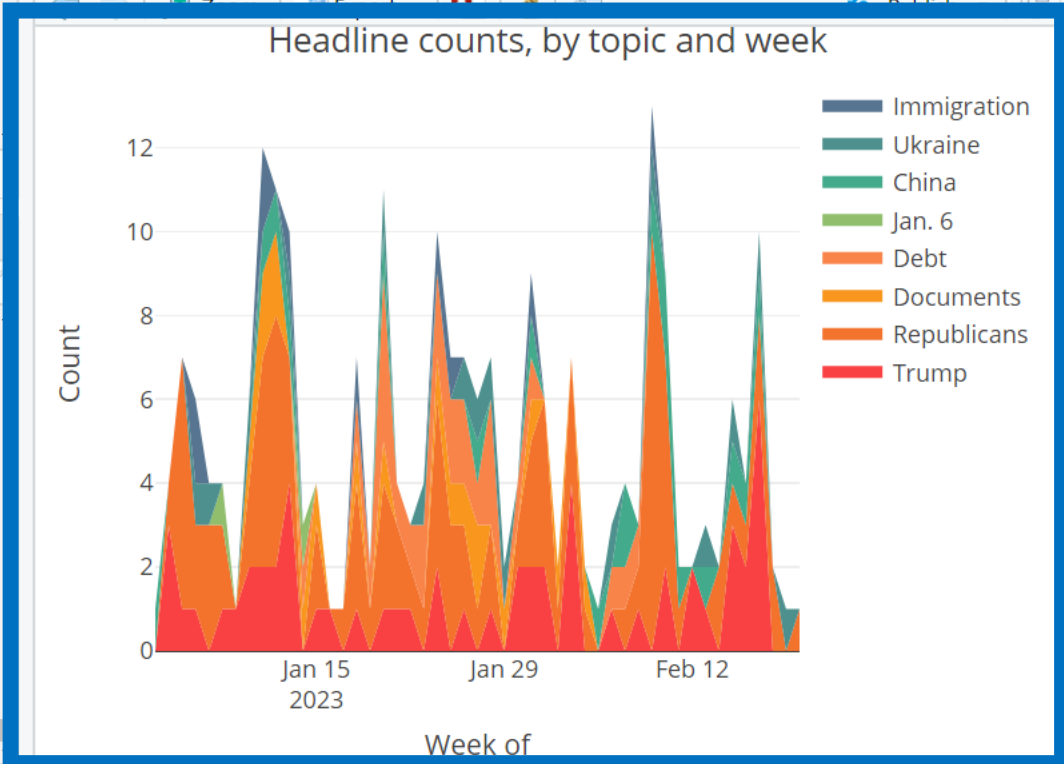
- AggByweek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- WordCounts 1726 obs. of 2 variables

Values

Hours	0.1
Minutes	4.1
searchterms	"border migra"

Files Plots Packages Help Viewer Presentation

A graph showing each topic's daily volume across the period searched. Plus ...



	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Showing 1 to 13 of 1,726 entries, 2 total columns

Console Terminal Background Jobs

... "mouse over"
 interactivity, and ...

```
> fig
> |
```

Environment History Connections Tutorial

422 MiB

R Global Environment

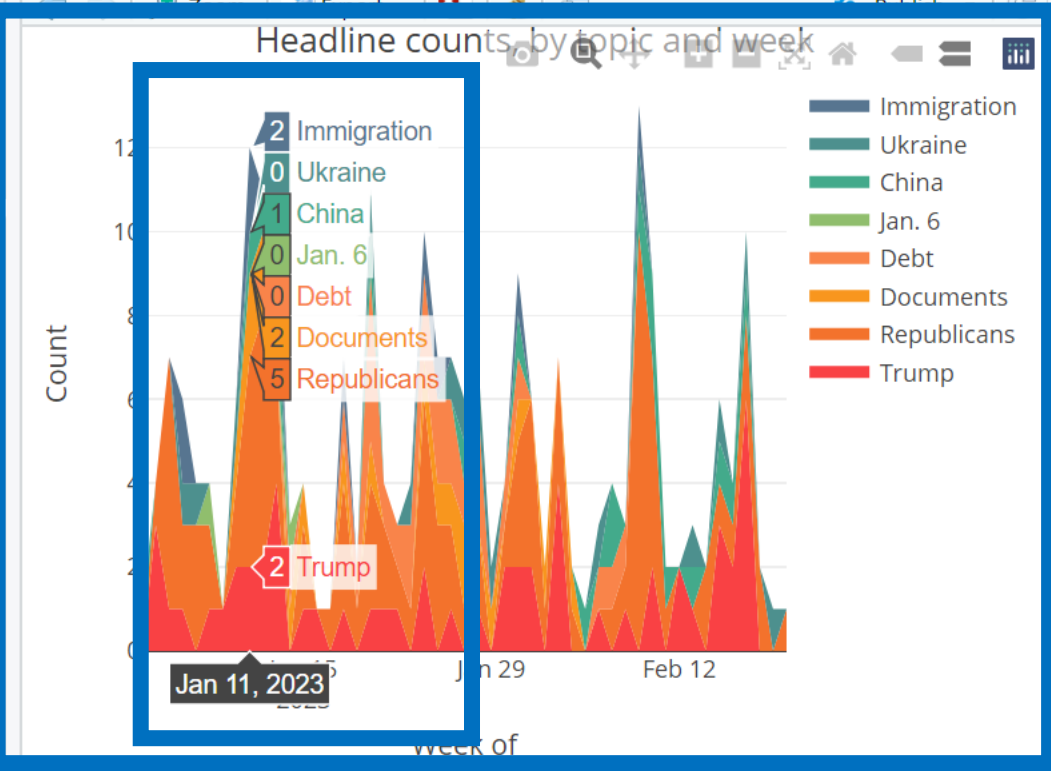
Data

- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- wordCounts 1726 obs. of 2 variables

Values

Hours 0.1

Files Plots Packages Help Viewer Presentation



	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Environment History Connections Tutorial

Global Environment 422 MiB

Data

- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- WordCounts 1726 obs. of 2 variables

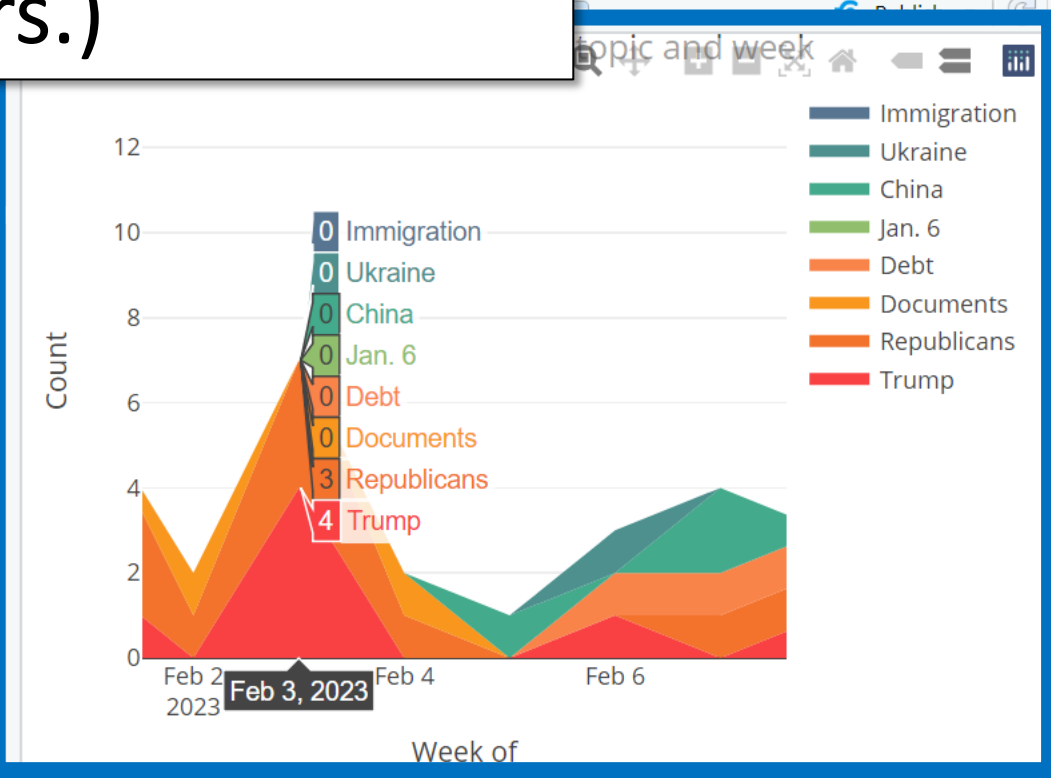
(And customizable colors.)

Console Terminal Background Jobs

```
> fig
```

... "mouse over" interactivity, and ...

... zoomability.



Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Showing 1 to 13 of 1,726 entries, 2 total columns

Environment History Connections Tutorial

Import Dataset 422 MiB

R Global Environment

Data

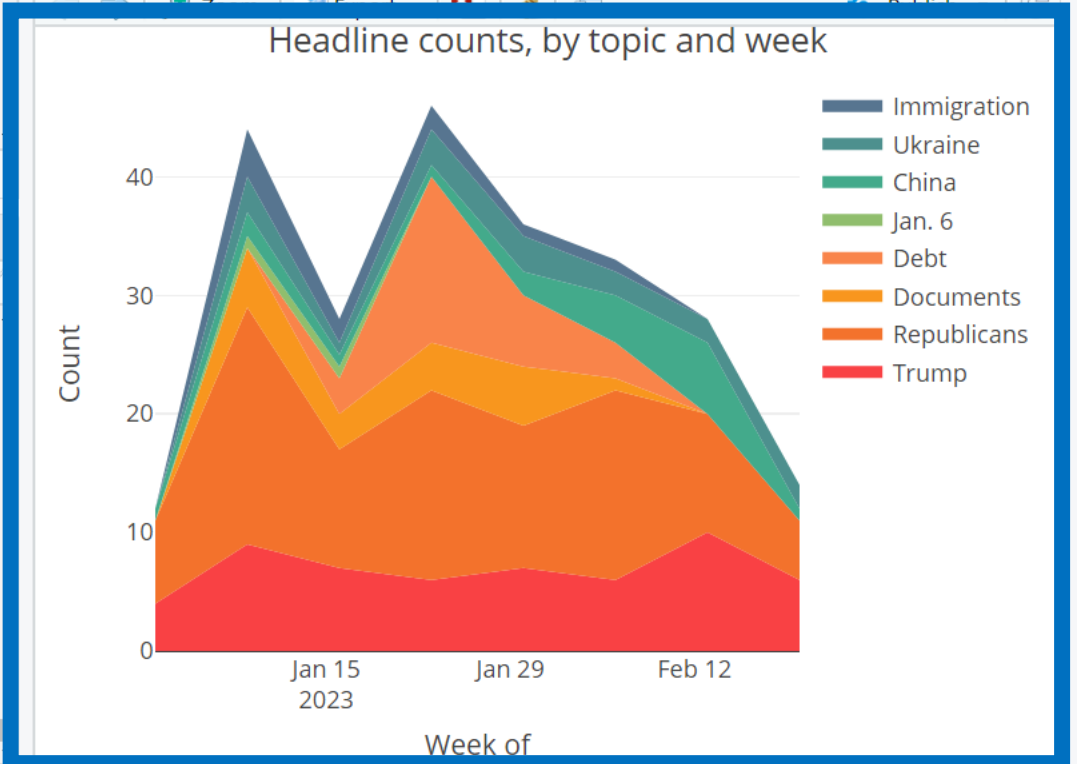
- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- wordCounts 1726 obs. of 2 variables

Values

Hours 0.1

Files Plots Packages Help Viewer Presentation

A graph of the same data, but by week ...



Console Terminal Background Jobs

```

>
> fig
>

```

Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

	Source	HeadlineCount
1	nytimes.com	139
2	washingtonpost.com	404

Showing 1 to 2 of 2 entries, 2 total columns

Environment History Connections Tutorial

Import Dataset 422 MiB

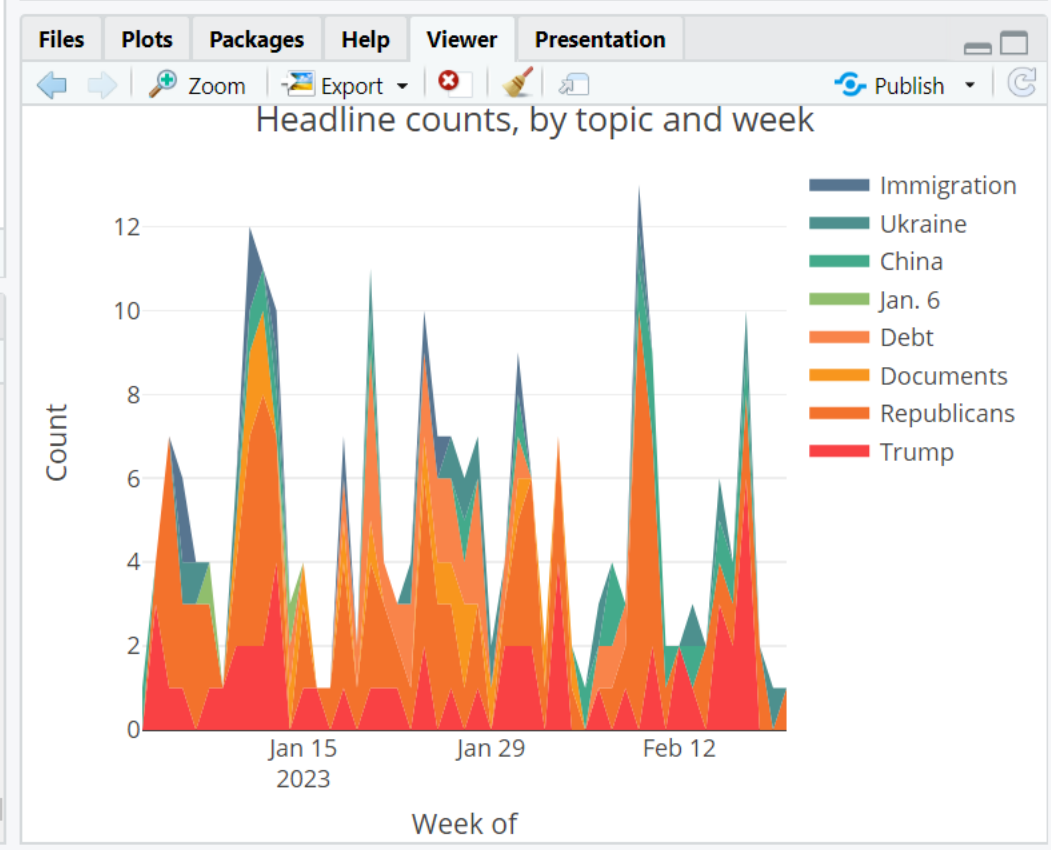
R Global Environment

Data

- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- wordCounts 1726 obs. of 2 variables

Values

Hours 0.1



Console Terminal Background Jobs

```
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
> fig <- fig %>% add_trace(y = ~Ukraine, name = 'Ukraine',
+                          fillcolor = '#4D908E')
+                          'Immigration',
+                          'by topic and week',
+                          'of',
+                          'SE)',
+                          '"',
+                          'E)')
+                          'of',
+                          'SE)',
+                          '"',
+                          'E)')
```

... the number of articles captured, by source ...

Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Showing 1 to 13 of 1,726 entries, 2 total columns

Environment History Connections Tutorial

Import Dataset 422 MiB

R Global Environment

Data

- AggByDay 49 obs. of 10 variables
- AggByWeek 8 obs. of 10 variables
- CountsBySour... 2 obs. of 2 variables
- fig List of 8
- Headlines 543 obs. of 15 variables
- WordCounts 1726 obs. of 2 variables

Values

Hours 0.1

Files Plots Packages Help Viewer Presentation

Folder Blank File Delete Rename

C:\Users\kblake\Desktop\R\Sandbox

Name	Size	Modified
..		
Sandbox.Rproj	218 B	Feb 25, 2023, 2:43 PM
'Joe Biden'20230102to20230219.csv	128.5 KB	Feb 25, 2023, 3:22 PM
GDELT scraper.R	9 KB	Feb 25, 2023, 2:58 PM

Console Terminal Background Jobs

```
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
> fig <- fig %>% add_trace(y = ~Ukraine, name = 'Ukraine',
+                          fillcolor = '#4D908E')
> fig <- fig %>% add_trace(y = ~Immigration, name = 'Immigration',
+                          fillcolor = '#577590')
> fig <- fig %>% layout(title = 'Headline counts, by topic',
+                       xaxis = list(title = "Week of",
+                                   showgrid = FALSE),
+                       yaxis = list(title = "Count",
+                                   showgrid = TRUE))
>
> fig
>
```

... and the data, exported as a .csv file automatically named after your search terms and date range.

Retrieve data for as many articles as you want, back to 2017 ...

... about any topic you can devise search terms for ...

... published in any indexed source, regardless of country or language ...

... for free.

	Title
15:00	New documents detail Sen . Ron John
00:00	The most intriguing revelations , new
30:00	Opinion Troops who refused covid va
15:00	Opinion Kevin McCarthy disastrous l
30:00	Missed some headlines over the holid
15:00	Opinion What George Santos has in
00:00	Time Is Running Out for Afghan Refug
00:00	The biggest health stories to watch in
00:00	Trump Tax Returns Are Only Part of th
00:00	Big Tech Is in Crisis . That Exactly Wha
00:00	Trump lawyers questioned Nevada 20.
00:00	Opinion Mikheil Saakashvili should n

Environment History Connections Tutorial

Import Dataset 422 MiB

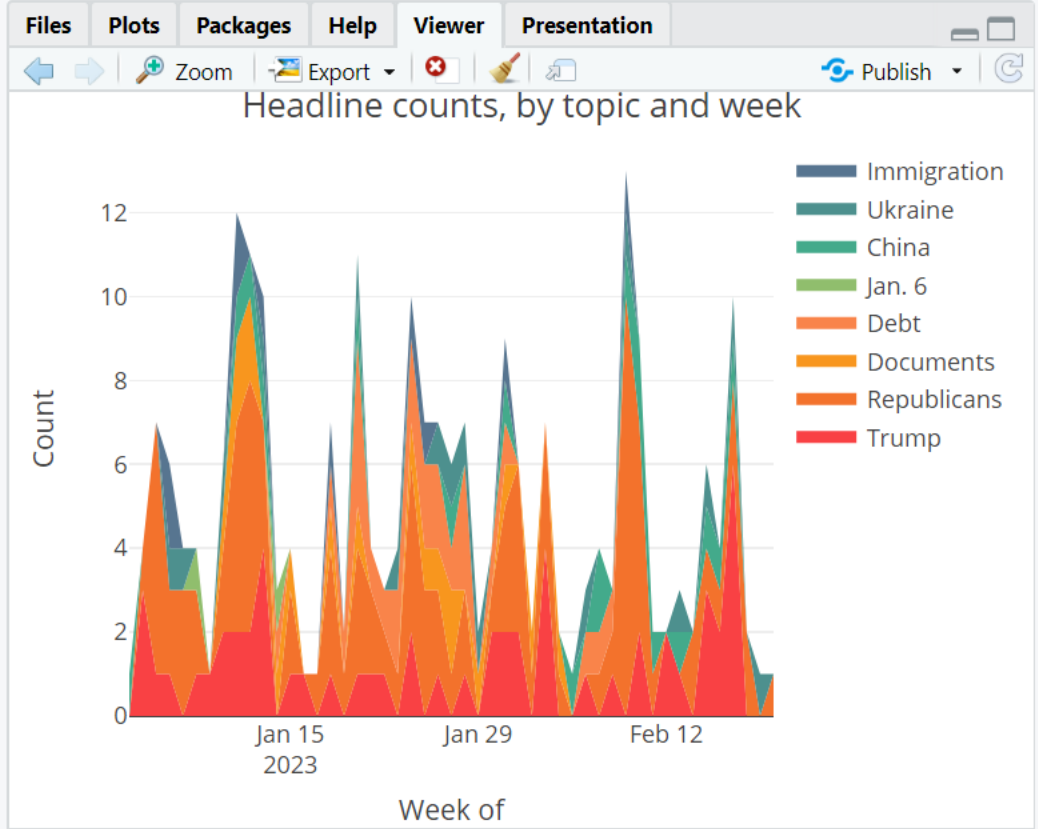
R Global Environment

Data

AggByDay	49 obs. of 10 variables
AggByWeek	8 obs. of 10 variables
CountsBySour...	2 obs. of 2 variables
fig	List of 8
Headlines	543 obs. of 15 variables
wordcounts	1726 obs. of 2 variables

Values

Hours	0.1
-------	-----



Data come from **GDELT**, the **G**lobal **D**atabase of **E**vents, **L**anguage and **T**one.

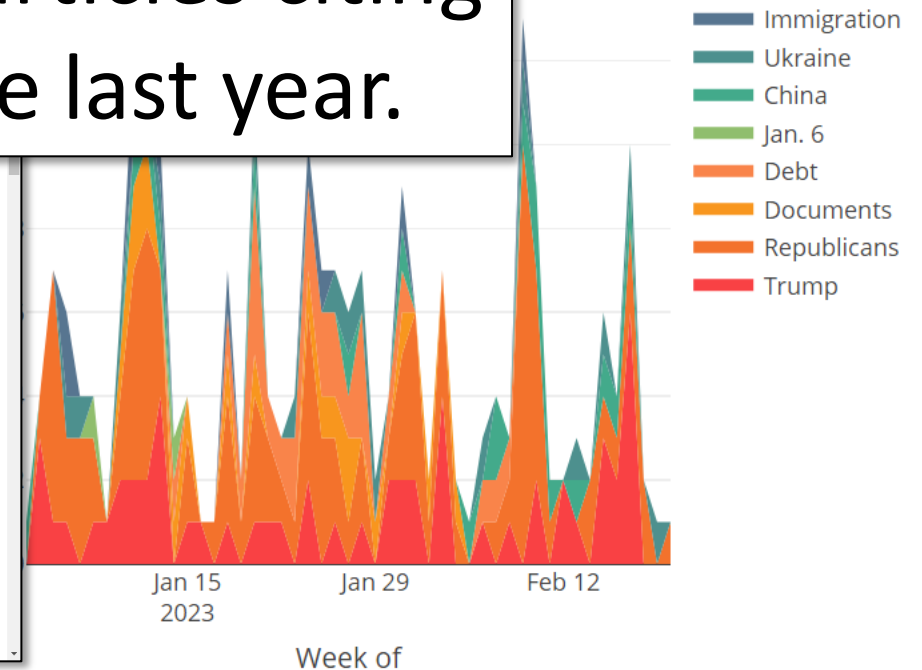
The GDELT Project

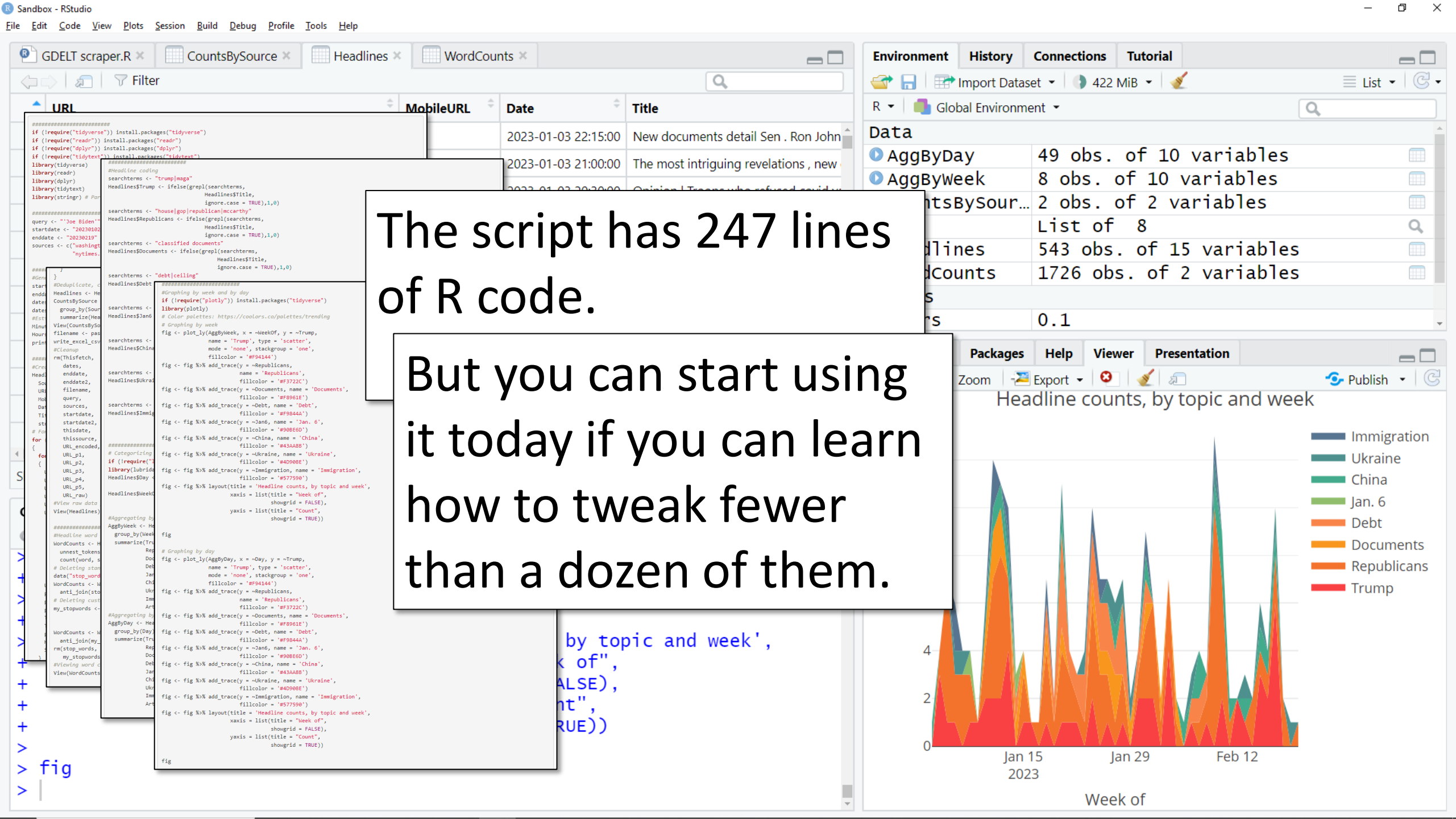
Google Scholar lists about 385 articles citing GDELT in the last year.

Google Scholar search results for "GDELT". The search bar shows "GDELT" and the results indicate "About 385 results (0.06 sec)". The results are sorted by date, showing articles added in the last year. The first article is "EU Climate Change News Index: Forecasting EU ETS prices with online news" by ÁD Hartvig and P Pálos, published in Finance Research Letters, 2023. The second article is "GTRL: An Entity Group-Aware Temporal Knowledge Graph Representation Learning Method" by X Tang and L Chen, an arXiv preprint from 2023. The third article is "Data innovations on protests in the United States" by C Dorff, G Adcox, and A Konet, published in the Journal of Peace Research, 2023.

RStudio Environment pane showing data objects. The objects listed are:

Object	Description
AggByDay	49 obs. of 10 variables
AggByWeek	8 obs. of 10 variables
CountsBySour...	2 obs. of 2 variables
fig	List of 8
Headlines	543 obs. of 15 variables
wordcounts	1726 obs. of 2 variables

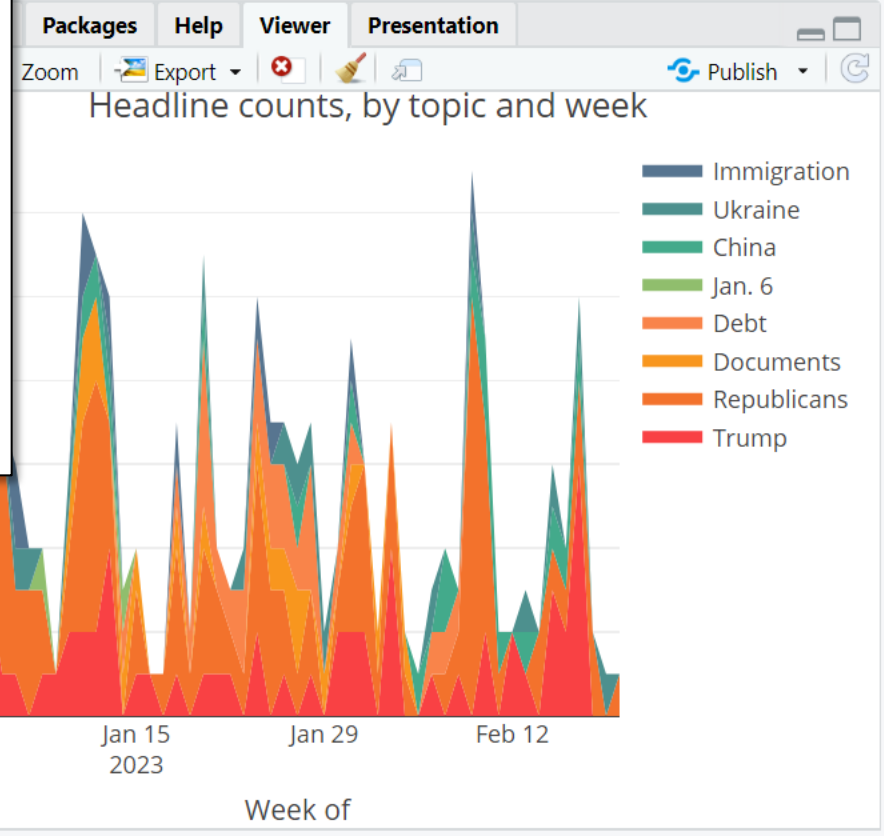




The script has 247 lines of R code. But you can start using it today if you can learn how to tweak fewer than a dozen of them.

by topic and week',
< of",
FALSE),
nt",
TRUE))

Environment	History	Connections	Tutorial
R	Global Environment	422 MiB	
Data			
AggByDay	49 obs. of 10 variables		
AggByWeek	8 obs. of 10 variables		
CountsBySource	2 obs. of 2 variables		
Lines	List of 8		
WordCounts	543 obs. of 15 variables		
Counts	1726 obs. of 2 variables		



```

query <- "'Joe Biden'" #Enter search term(s)
startdate <- "20230102" #Enter preferred start date
enddate <- "20230219" #Enter preferred end date
sources <- c("washingtonpost.com",
            "nytimes.com") #Enter sources to search

```

22 MiB | [brush icon]

List [dropdown]

[search icon]

- . of 10 variables [calendar icon]
- of 10 variables [calendar icon]
- of 2 variables [calendar icon]
- f 8 [search icon]
- s. of 15 variables [calendar icon]
- obs. of 2 variables [calendar icon]

The “query” line is where you specify your search terms.

query <- "'Biden'" : The word “Biden”

query <- "'Joe Biden'" : The **phrase** “Joe Biden”

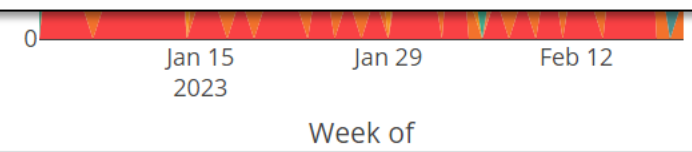
query <- "'Joe Biden' 'Jill Biden'" : Both “Joe Biden” **and** “Jill Biden”

query <- " ('Joe Biden' OR 'Jill Biden') " : **Either** “Joe Biden” **or** “Jill Biden,” or both

```

>
> fig
>

```



GDELT Headline Fetch

Ken Blake

2023-02-24

Intro

This R script will use the [GDELT 2.0 DOC API](#) to fetch individual article URLs, headlines and publication dates for online one or more user-specified keywords or phrases in their body copy and were published by one or more user-specified user-specified time period. The GDELT API offers data for stories published in January 2017 or later.

Additionally, the script can export the fetched data in comma-separated-value format, produce a sorted word frequency code user-specified keywords as either present (1) or absent (0) in each headline, aggregate the data to produce weekly headlines that mention each issue, and graph these weekly and daily counts as interactive stacked-area charts.

Questions about this script may be directed to Dr. Ken Blake. See: <https://drkblake.com/>. Scroll to the end of this page for sections immediately below divide the script into code chunks and offer instructions and explanations.

Required packages

The script requires several R packages. This first batch is needed for the script's basic retrieval, export, and word any packages that haven't already been installed and ensures that the requisite package libraries are activated for

```
#####  
if (!require("tidyverse")) install.packages("tidyverse")  
if (!require("readr")) install.packages("readr")  
if (!require("dplyr")) install.packages("dplyr")  
if (!require("tidytext")) install.packages("tidytext")  
library(tidyverse)  
library(readr)  
library(dplyr)  
library(tidytext)  
library(stringr) # Part of the tidyverse package
```

Query setup

The next block of code is where you tell the script which terms to look for, which news outlets to search, and the words, multi-word phrases, and combinations of the two can be searched. Here are some example queries and the each will search stories for:

```
query <- "'Biden'" : The word "Biden"
```

```
query <- "'Joe Biden'" : The phrase "Joe Biden"
```

```
query <- "'Joe Biden' 'Jill Biden'" : Both "Joe Biden" and "Jill Biden"
```

```
query <- "'(Joe Biden' OR 'Jill Biden)'" : Either "Joe Biden" or "Jill Biden," or both
```

Punctuation is critical in all of these query variations. Be sure each double apostrophe, single apostrophe and, if present, parentheses, is included and placed correctly.

This example sets up a search of two news outlets for all stories that mention the phrase "Joe Biden" and that were published between Jan. 2 and Feb. 19 of 2023.

Incidentally, all hints I'll go over are covered in the script's companion web page.

See:

<https://rpubs.com/drkblake/1007551>

Also there: A full copy of the script.

You can have a copy of this PowerPoint, too.

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Environment History Connections Tutorial

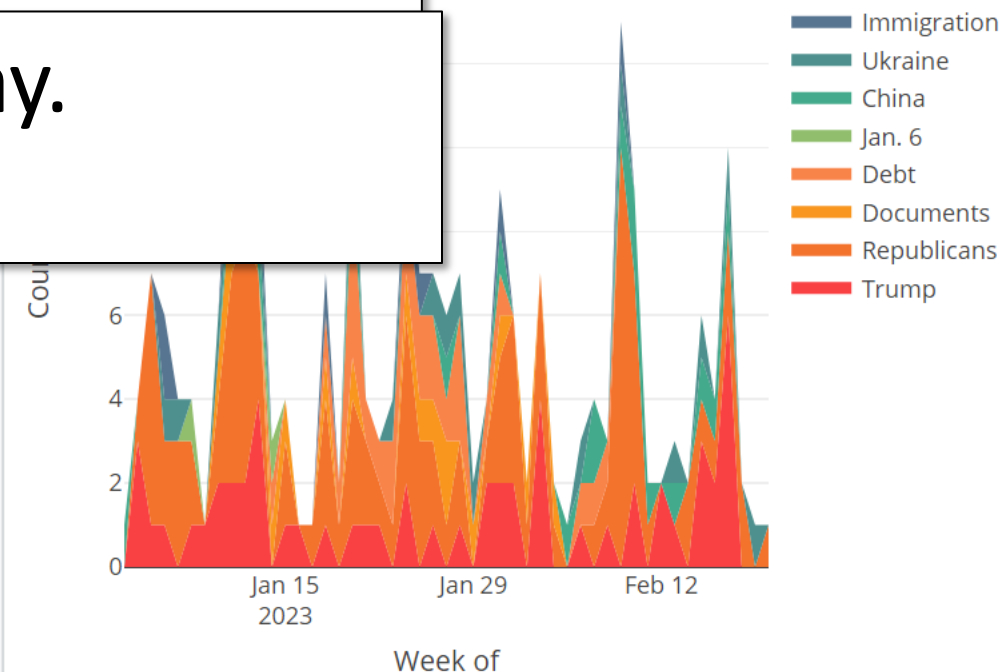
```
query <- "'Joe Biden'" #Enter search term(s)
startdate <- "20230102" #Enter preferred start date
enddate <- "20230219" #Enter preferred end date
sources <- c("washingtonpost.com",
            "nytimes.com") #Enter sources to search
```

Specify a “startdate” for your search in YYYYMMDD format.

Specify an “enddate” in the same way.

Con

```
> plot(figure = fig, titlecolor = "#4D908E")
> fig <- fig %>% add_trace(y = ~Immigration, name = 'Immigration',
+                         fillcolor = '#577590')
> fig <- fig %>% layout(title = 'Headline counts, by topic and week',
+                       xaxis = list(title = "week of",
+                                   showgrid = FALSE),
+                       yaxis = list(title = "Count",
+                                   showgrid = TRUE))
> fig
```



GDEL scraper.R x CountsBySource x Headlines x WordCounts x

Environment History Connections Tutorial

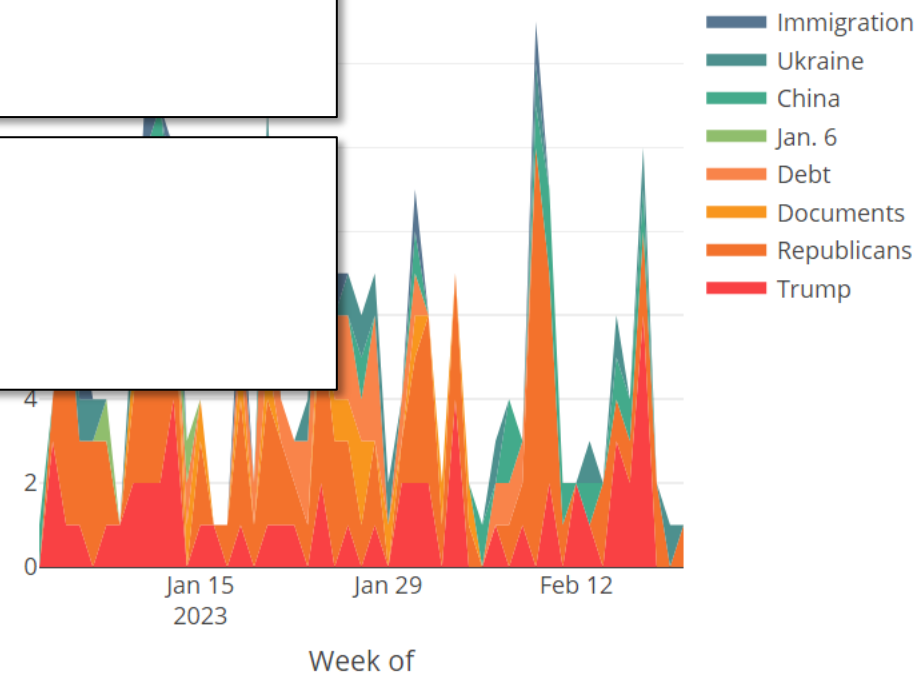
```
query <- "'Joe Biden'" #Enter search term(s)
startdate <- "20230102" #Enter preferred start date
enddate <- "20230219" #Enter preferred end date
sources <- c("washingtonpost.com",
            "nytimes.com") #Enter sources to search
```

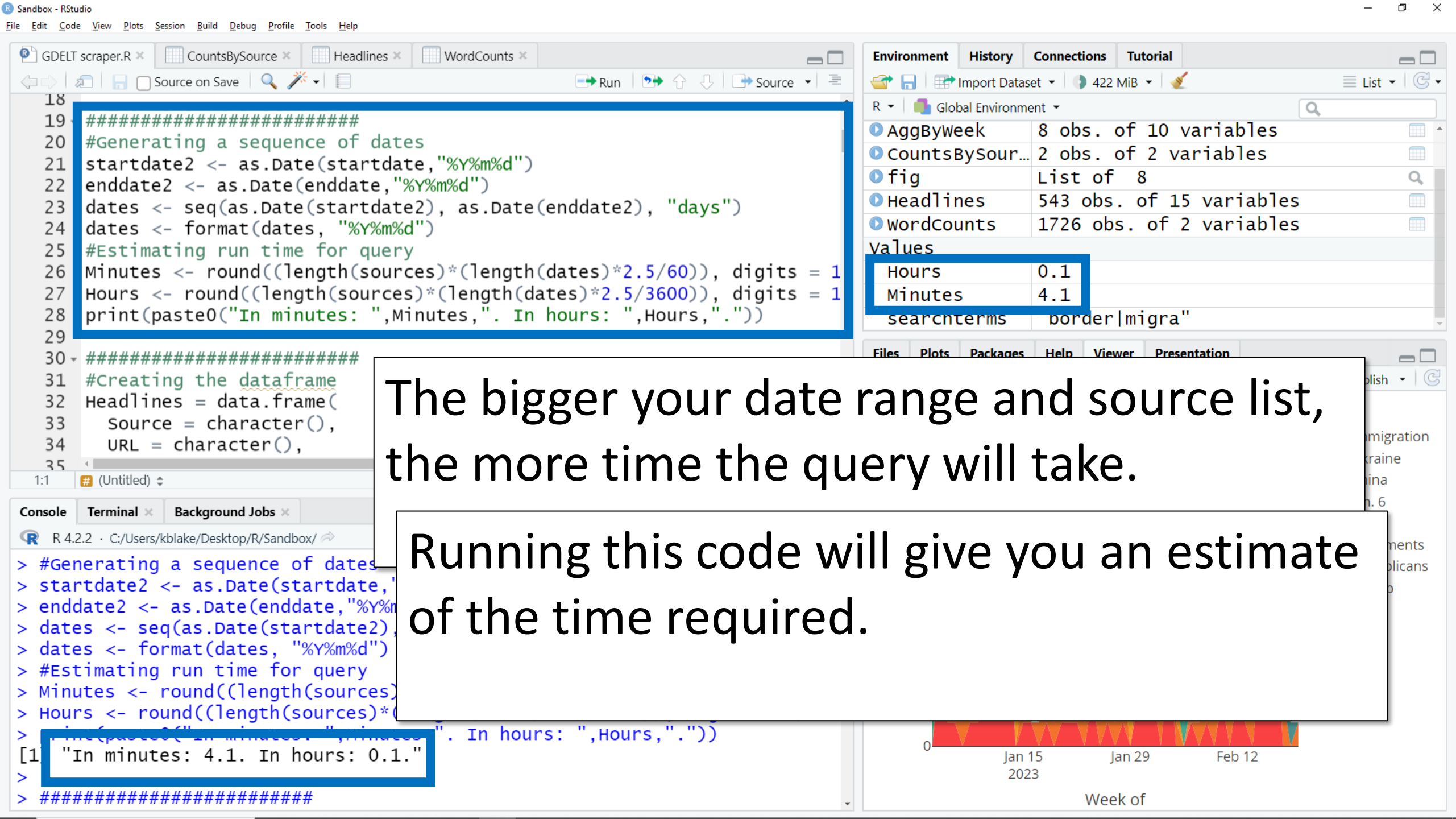
Add as many sources as you like, using the root of the source's web address.

Punctuation matters ... a lot.

```
xaxis = list(title = "week of",
            showgrid = FALSE),
yaxis = list(title = "Count",
            showgrid = TRUE))
```

```
> fig
```





The bigger your date range and source list, the more time the query will take.

Running this code will give you an estimate of the time required.

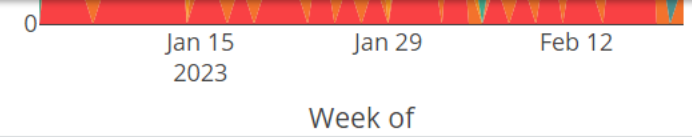
```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source
29
30 #####
31 #Creating the dataframe
32 Headlines = data.frame(
33   source = character(),
34   URL = character(),
35   MobileURL = character(),
36   Date = character(),
37   Title = character(),
38   stringsAsFactors = FALSE)
39 # For loops
40 for (thissource in sources)
41 {
42   for (thisdate in dates)
43   {
44     URL_p1 <- "https://api.gdeftproject.org/api/v2/doc/doc?query="
45     URL_p2 <- " domainis:"
46
```

This code starts a “loop” that sends the GDELT API a request – one at a time – for each search terms/source/date/combination.

```
Console Terminal Background Jobs
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] "Getting data for"
[1] "20230219 washingtonpost.com"
[1] "https://api.gdeftproject.org/api/v2/doc/doc?query='Joe%20Biden'%20domainis:washingtonpost.com&mode=artlist&maxrecords=250&sort=datedesc&startdatetime=20230219000000&enddatetime=20230219235959&format=CSV"
[1] "5 rows"
[1] "Getting data for"
[1] "20230102 nytimes.com"
[1] "https://api.gdeftproject.org/api/v2/doc/doc?query='Joe%20Biden'%20domainis:nytimes.com&mode=artlist&maxrecords=250&sort=datedesc&startdatetime=20230102000000&enddatetime=20230102235959&format=CSV"
[1] "2 rows"
```

Each result shows here as the loop runs.



```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Source on Save Run Source

29
30 #####
31 #Creating the dataframe
32 Headlines = data.frame(
33   source = character(),
34   URL = character(),
35   MobileURL = character(),
36   Date = character(),
37   Title = character(),
38   stringsAsFactors = FALSE)
39 # For loops
40 for (thissource in sources)
41 {
42   for (thisdate in dates)
43   {
44     URL_p1 <- "https://api.gdeftproject.org/api/v2/doc/doc?query="
45     URL_p2 <- " domainis:"
46
```

Environment History Connections Tutorial

422 MiB

R Global Environment

Data

AggByDay	49 obs. of 10 variables
AggByWeek	8 obs. of 10 variables
CountsBySour...	2 obs. of 2 variables
fig	List of 8
Headlines	543 obs. of 15 variables
WordCounts	1720 obs. of 2 variables

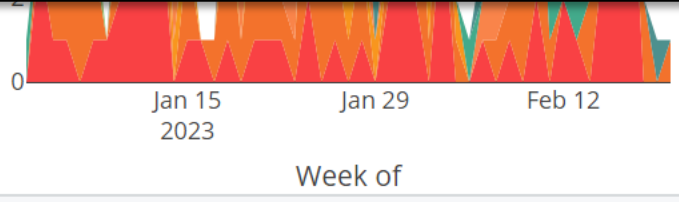
Each iteration adds data to the "Headlines" data frame.

Console Terminal Background Jobs

R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/

```
[1] "Getting data for"
[1] "20230219 washingtonpost.com"
[1] "https://api.gdeftproject.org/api/v2/doc/doc?query='Joe%20Biden'%20domainis:washingtonpost.com&mode=artlist&maxrecords=250&sort=datedesc&startdatetime=20230219000000&enddatetime=20230219235959&format=CSV"
[1] "5 rows"
[1] "Getting data for"
[1] "20230102 nytimes.com"
[1] "https://api.gdeftproject.org/api/v2/doc/doc?query='Joe%20Biden'%20domainis:nytimes.com&mode=artlist&maxrecords=250&sort=datedesc&startdatetime=20230102000000&enddatetime=20230102235959&format=CSV"
[1] "2 rows"
```

Stop the loop, and "Headlines" will contain all data collected so far.



```
GDELT scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source
69 #Deduplicate, check, and export data
70 Headlines <- Headlines[!duplicated(Headlines$URL),]
71 countsBySource <- Headlines %>%
72   group_by(Source) %>%
73   summarize(HeadlineCount = n())
74 view(CountsBySource)
75 filename <- paste0(query,startdate,"to",enddate,".csv")
76 write_excel_csv(Headlines,filename)
77 #Cleanup
78 rm(Thisfetch,
79   dates,
80   enddate,
81   enddate2,
82   filename,
83   query,
84   sources,
85   startdate
```

After the loop, a built-in process deletes any duplicate records.

```
Console Terminal x Background Jobs x
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] U R OWS
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> countsBySource <- Headlines %>%
+   group_by(Source) %>%
+   summarize(HeadlineCount = n())
> view(CountsBySource)
> filename <- paste0(query,startdate,"to",enddate,".csv")
> write_excel_csv(Headlines,filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate
```

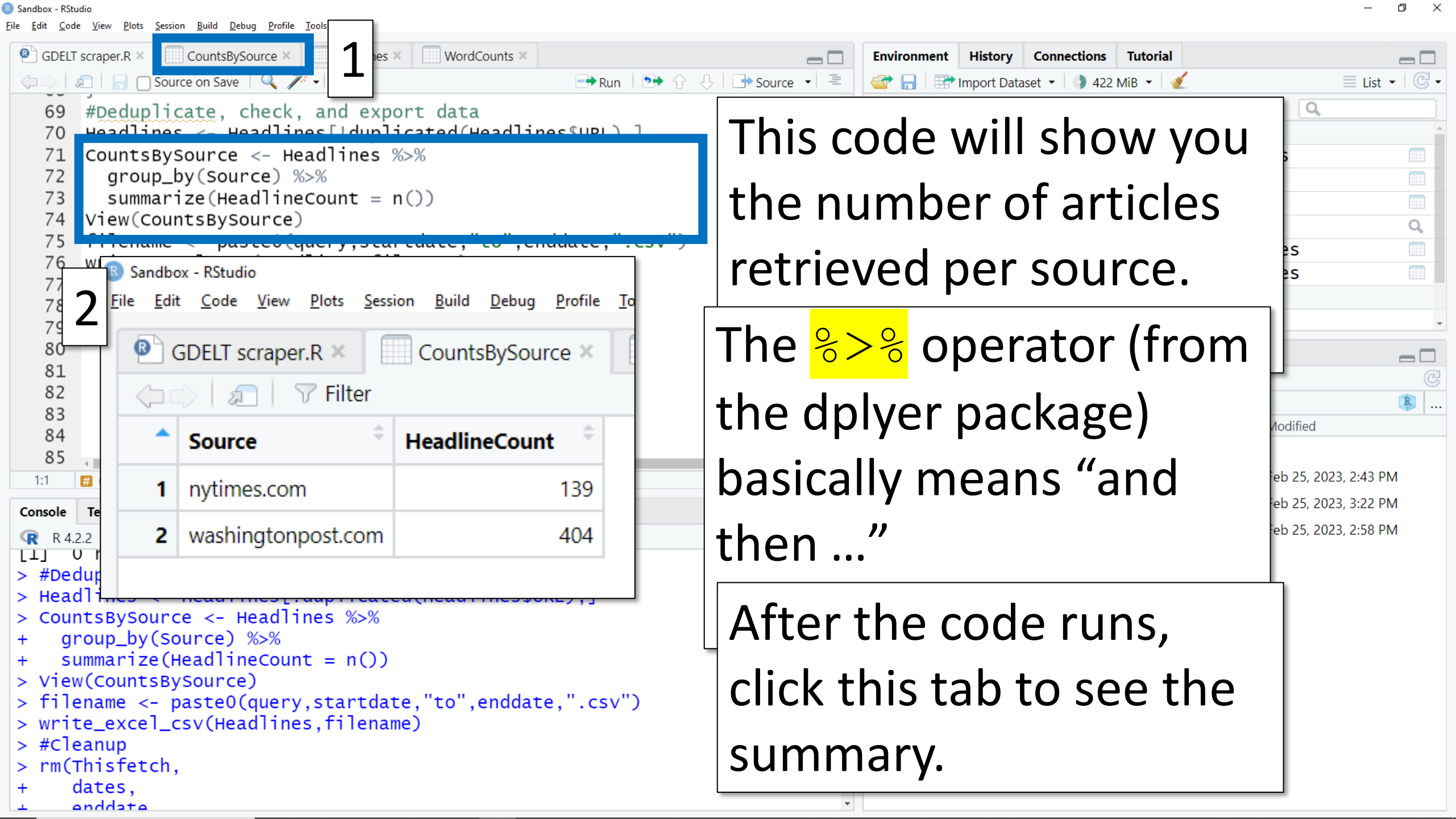
Environment History Connections Tutorial
Import Dataset 422 MiB

Hours 0.1

Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename

C > Users > kblake > Desktop > R > Sandbox

	Name	Size	Modified
<input type="checkbox"/>	..		
<input type="checkbox"/>	Sandbox.Rproj	218 B	Feb 25, 2023, 2:43 PM
<input type="checkbox"/>	'Joe Biden'20230102to20230219.csv	128.5 KB	Feb 25, 2023, 3:22 PM
<input type="checkbox"/>	GDELT scraper.R	9 KB	Feb 25, 2023, 2:58 PM



1

```
CountsBySource <- Headlines %>%  
  group_by(Source) %>%  
  summarize(HeadlineCount = n())  
View(CountsBySource)
```

This code will show you the number of articles retrieved per source.

2

	Source	HeadlineCount
1	nytimes.com	139
2	washingtonpost.com	404

The `%>%` operator (from the dplyer package) basically means “and then ...”

After the code runs, click this tab to see the summary.

```
> #Deduplicate, check, and export data  
> Headlines <- Headlines[!duplicated(Headlines$URL), ]  
> CountsBySource <- Headlines %>%  
+   group_by(Source) %>%  
+   summarize(HeadlineCount = n())  
> View(CountsBySource)  
> filename <- paste0(query,startdate,"to",enddate,".csv")  
> write_excel_csv(Headlines,filename)  
> #Cleanup  
> rm(Thisfetch,  
+   dates,  
+   enddate
```



```

99 #####
100 #Headline word counts
101 WordCounts <- Headlines %>%
102   unnest_tokens(word,Title) %>%
103   count(word, sort = TRUE)
104 # Deleting standard stop words
105 data("stop_words")
106 WordCounts <- WordCounts %>%
107   anti_join(stop_words)
108 # Deleting custom stop words
109 my_stopwords <- tibble(word = c("and",
110                                "the",
111                                "etc."))
112 WordCounts <- WordCounts %>%
113   anti_join(my_stopwords)
114 rm(stop_words,
115     mv stopwords)

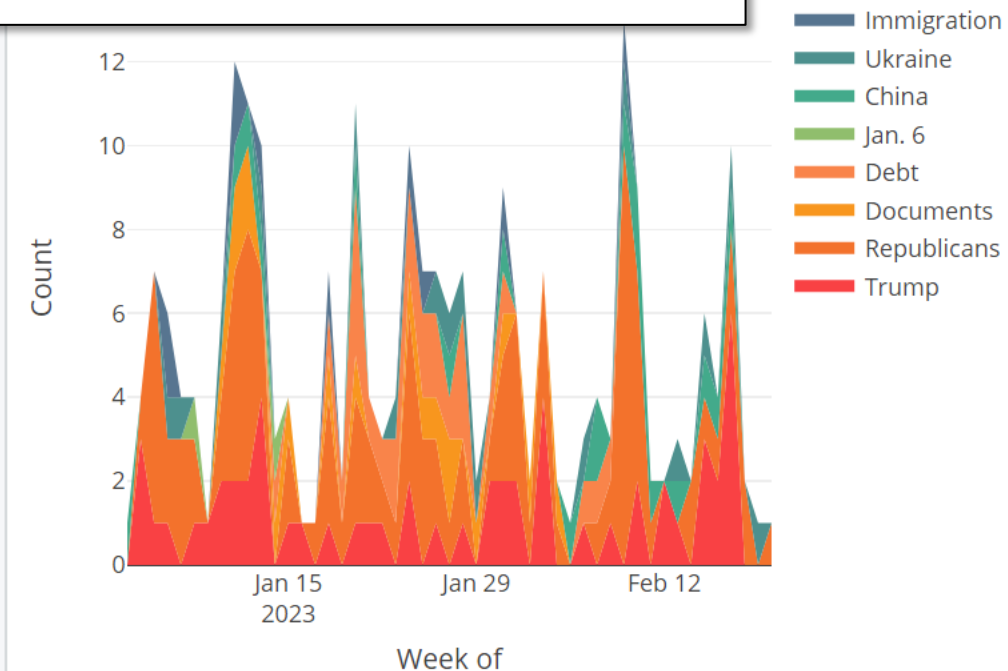
```

This code deletes common “stop words” from the headline word count.

```

R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] 0 rows
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> CountsBySource <- Headlines %>%
+   group_by(source) %>%
+   summarize(HeadlineCount = n())
> view(CountsBySource)
> filename <- paste0(query,startdate,"to",enddate,".csv")
> write_excel_csv(Headlines,filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate

```



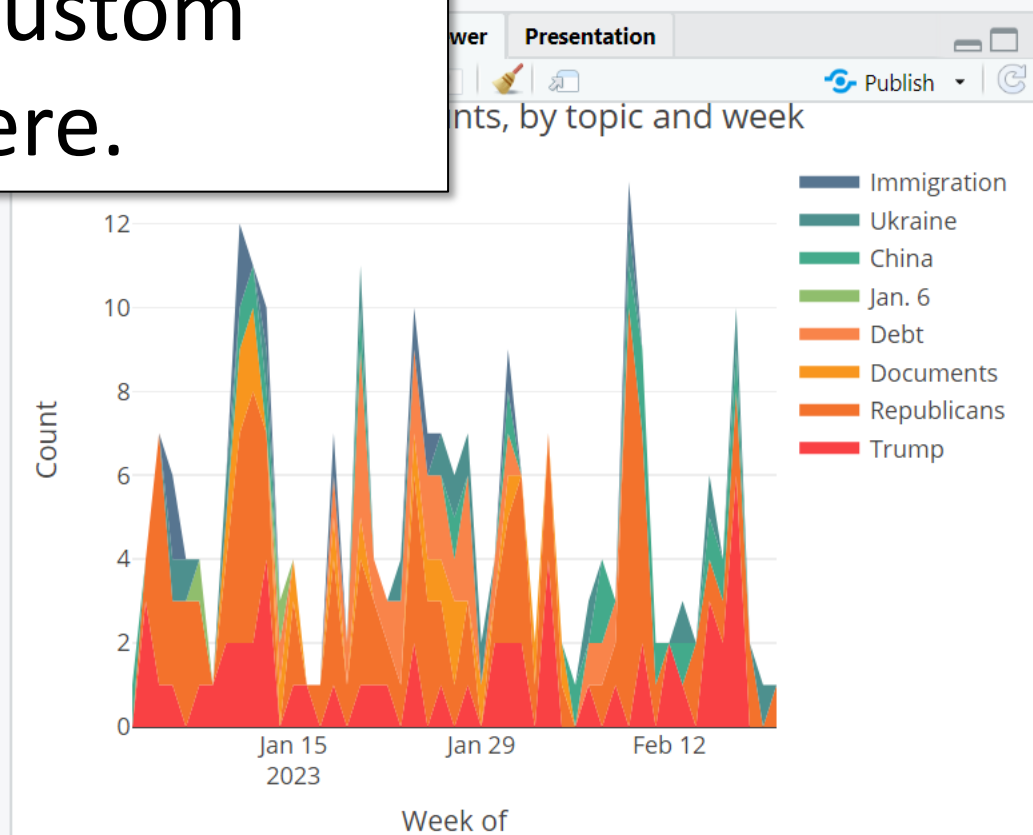
```
GDELTL scraper.R x CountsBySource x Headlines x WordCounts x  
Source on Save Run Source  
99 #####  
100 #Headline word counts  
101 wordCounts <- Headlines %>%  
102   unnest_tokens(word,Title) %>%  
103   count(word, sort = TRUE)  
104 # Deleting standard stop words  
105 data("stop_words")  
106 wordCounts <- wordCounts %>%  
107   anti_join(stop_words)  
108 # Deleting custom stop words  
109 my_stopwords <- tibble(word = c("and",  
110   "the",  
111   "etc."))  
112 wordCounts <- wordCounts %>%  
113   anti_join(my_stopwords)  
114 rm(stop_words,  
115   mv stopwords)  
116  
1:1 # (Untitled) R Script
```

```
c("and",  
  "the",  
  "etc.")
```

You can add custom
stop words here.

Environment	History	Connections	Tutorial
R	Global Environment	422 MiB	
Data			
AggByDay	49 obs. of 10 variables		
AggByweek	8 obs. of 10 variables		
CountsBySour...	2 obs. of 2 variables		
fig	List of 8		
Headlines	543 obs. of 15 variables		
wordCounts	1726 obs. of 2 variables		

```
Console Terminal x Background Jobs x  
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/  
[1] 0 rows  
> #Deduplicate, check, and export data  
> Headlines <- Headlines[!duplicated(Headlines$URL),]  
> countsBySource <- Headlines %>%  
+   group_by(source) %>%  
+   summarize(HeadlineCount = n())  
> view(countsBySource)  
> filename <- paste0(query,startdate,"to",enddate,".csv")  
> write_excel_csv(Headlines,filename)  
> #Cleanup  
> rm(Thisfetch,  
+   dates,  
+   enddate
```

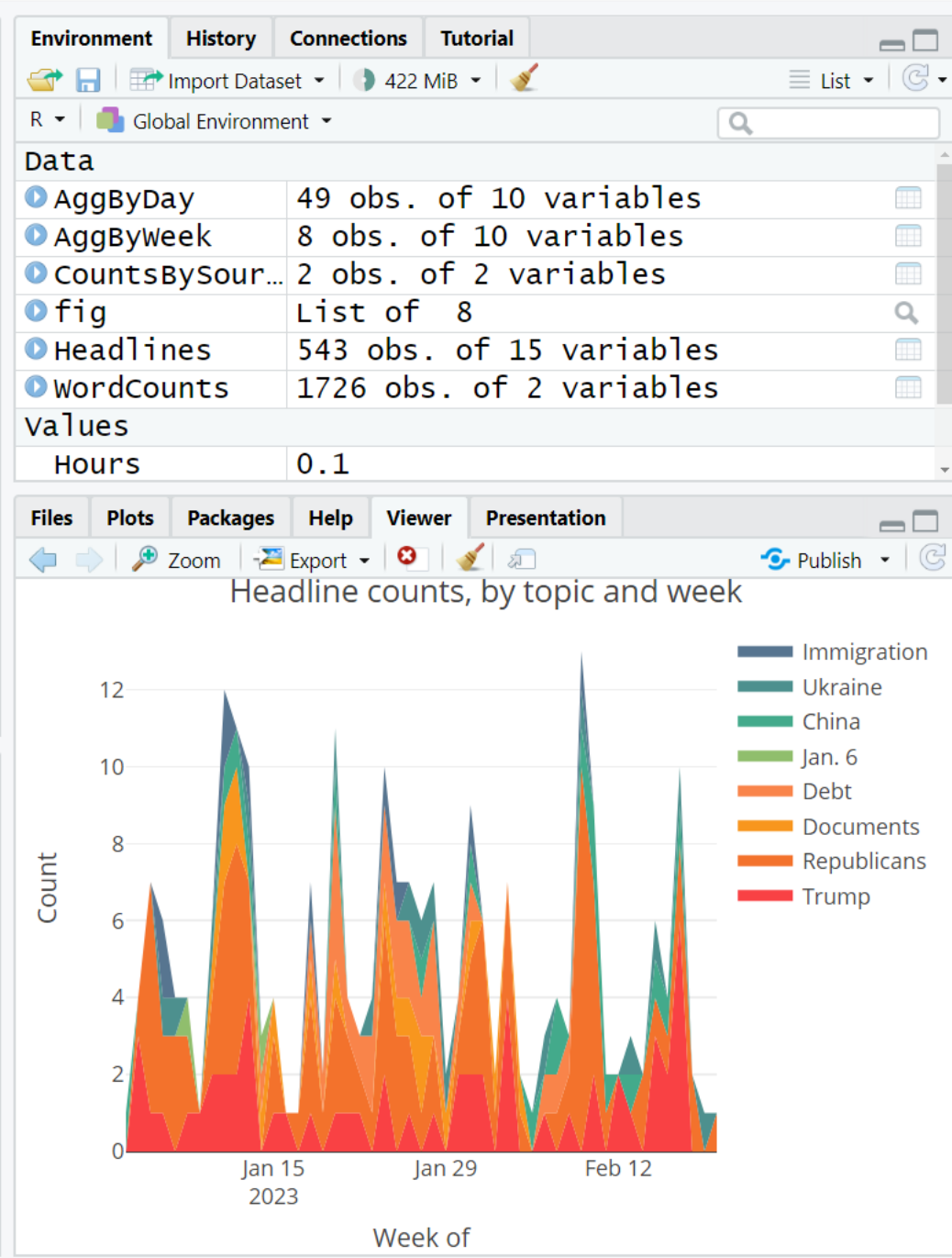


	word	n
1	biden	131
2	opinion	93
3	post	83
4	washington	82
5	trump	48
6	house	39
7	gop	34
8	documents	27
9	debt	26
10	republicans	24
11	classified	23
12	6	19

Showing 1 to 13 of 1,726 entries, 2 total columns

Sorted headline word counts are under this tab, and they look like this.

```
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] U R OWS
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> CountsBySource <- Headlines %>%
+   group_by(source) %>%
+   summarize(HeadlineCount = n())
> view(CountsBySource)
> filename <- paste0(query,startdate,"to",enddate,".csv")
> write_excel_csv(Headlines,filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate
```



Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

1 Filter 2

Date	Title	Source	Trump	Re
2023-01-03 22:15:00	New documents detail Sen . Ron Johnson asking about elect...	washingtonpost.com	0	
2023-01-10 00:45:00	Classified documents found in Biden post - VP office under ...	washingtonpost.com	0	
2023-01-11 23:45:00	Second Biden search yields additional classified documents	washingtonpost.com	0	
2023-01-11 09:30:00	Biden , Trump and classified documents : An explainer	washingtonpost.com	1	
2023-01-11 00:45:00	Biden surprised by classified documents as Hill demands m...	washingtonpost.com	0	
2023-01-12 21:15:00	Who is Robert Hur , special counsel in Biden documents cas...	washingtonpost.com	0	
2023-01-13 23:15:00	Biden classified documents are a media test and opportunity	washingtonpost.com	0	
2023-01-14 17:30:00	Lawyers found more classified documents at Joe Biden home	washingtonpost.com	0	
2023-01-17 19:15:00	The impressively weak effort to whatabout Biden classified ...	washingtonpost.com	0	
2023-01-17 13:00:00	The Difference Is That Biden Gave the Documents Back	washingtonpost.com	0	
2023-01-22 01:45:00	FBI searched Biden home , found documents marked classifi...	washingtonpost.com	0	

Showing 1 to 11 of 27 entries, 15 total columns (filtered from 543 total entries)

On the Headlines tab, you can filter the headlines by keyword to explore how each keyword is used.

Here, you see that many – but not all – headlines mentioning “documents” are about Biden’s “classified documents.”

```
Console Terminal Background Jobs
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] 0 rows
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> CountsBySource <- Headlines %>%
+   group_by(source) %>%
+   summarize(HeadlineCount = n())
> view(CountsBySource)
> filename <- paste0(query,startdate,"to",enddate,".csv")
> write_excel_csv(Headlines,filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate
```

Sandbox - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x

Filter

Date	Title	Source	Trump	Republicans	Do
2023-01-03 21:00:00	The most intriguing revelations , new evidence from Jan . 6 t...	washingtonpost.com	0	0	0
2023-01-04 12:15:00	On official tours of the U . S . Capitol there is no mention of ...	washingtonpost.com	0	0	0
2023-01-04 10:00:00	Jan . 6 committee yet again debunks Trump claim of 10,000...	washingtonpost.com	1	0	0
2023-01-05 18:30:00	Biden to award Presidential Citizens Medal to 1	washingtonpost.com	0	0	0
2023-01-06 21:15:00	Biden warns of dangers posed by 'big lie', reward	washingtonpost.com	0	0	0
2023-01-06 12:00:00	2 years after Jan . 6 , speaker scrap paralyzes Congress again	washingtonpost.com	0	0	0
2023-01-06 11:45:00	Jan . 6 rioters cast as patriot by supporters who have raised...	washingtonpost.com	0	0	0
2023-01-07 19:30:00	The irony of the McCarthy speaker bid drama playing out o...	washingtonpost.com	0	1	0
2023-01-07 14:15:00	Ashli Babbitt : From Jan . 6 Capitol rioter to Trump - embrac...	washingtonpost.com	1	0	0
2023-01-09 20:30:00	Brazil attack from Bolsonaro supporters shows similarities to...	washingtonpost.com	0	0	0
2023-01-12 17:15:00	Proud Boys trial : Live coverage and latest updates , January ...	washingtonpost.com	0	0	0

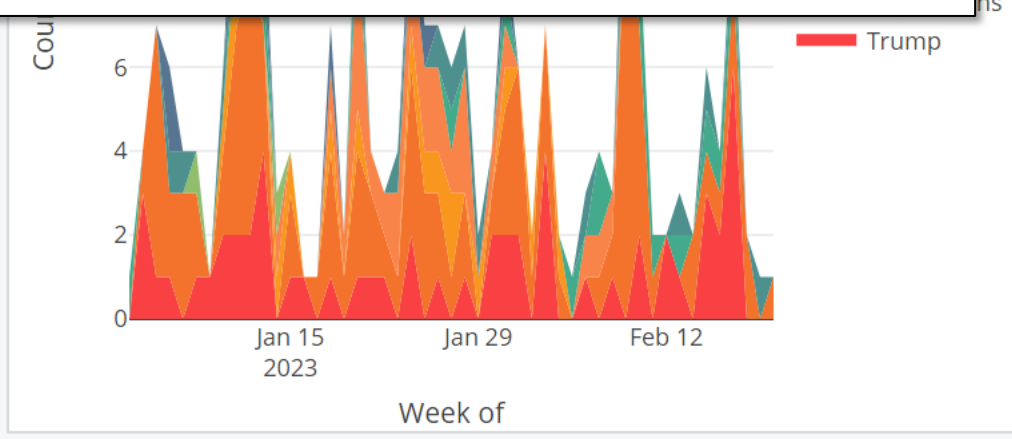
Showing 1 to 11 of 20 entries, 15 total columns (filtered from 543 total entries)

The same method shows that "Jan. 6" will flag headlines about the Jan. 6 capitol riot.

A mouse hover will show you the full text of any cell.

Console Terminal Background Jobs

```
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] 0 rows
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> CountsBySource <- Headlines %>%
+   group_by(source) %>%
+   summarize(HeadlineCount = n())
> view(CountsBySource)
> filename <- paste0(query,startdate,"to",enddate,".csv")
> write_excel_csv(Headlines,filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate
```



```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source

118
119 #####
120 #headline coding
121 searchterms <- "trump|maga"
122 Headlines$Trump <- ifelse(grepl(searchterms,
123                               Headlines$title,
124                               ignore.case = TRUE),1,0)
125 searchterms <- "house|gop|republican|mccarthy"
126 Headlines$Republicans <- ifelse(grepl(searchterms,
127                                       Headlines$title,
```

You can then customize this code to categorize headlines by keywords.

Here, a headline gets a "1" if it mentions "trump" or "maga" & a "0" if it doesn't.

Trump	Republicans	Documents	Debt	Jan6	China	Ukraine	Immigration	Day	WeekOf
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
1	1	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
1	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-03	2023-01-02
1	0	0	0	0	0	0	0	2023-01-05	2023-01-02
0	0	0	0	0	0	0	0	2023-01-05	2023-01-02

The codes end up stored in a dataset column called "Trump."

```
> #Cleanup
> rm(Thisfetch,
+     dates,
+     enddate
```

Week of

```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDEL scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source

118
119 #####
120 #headline coding
121 searchterms <- "trump|maga"
122 Headlines$Trump <- ifelse(grepl(searchterms,
123                               Headlines$title,
124                               ignore.case = TRUE), 1, 0)
125 searchterms <- "house|gop|republican|mccarthy"
126 Headlines$Republicans <- ifelse(grepl(searchterms,
127                                       Headlines$title,
```

“Or,” represented by a “|” character, is the only Boolean operator available, here.

Trump	Republicans	Documents	Debt	Jan6	China	Ukraine	Immigration	Day	WeekOf
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
1	1	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-04	2023-01-02
1	0	0	0	0	0	0	0	2023-01-04	2023-01-02
0	0	0	0	0	0	0	0	2023-01-03	2023-01-02
1	0	0	0	0	0	0	0	2023-01-05	2023-01-02
0	0	0	0	0	0	0	0	2023-01-05	2023-01-02

Showing 1 to 12 of 543 entries, 15 total columns

But the search is string-based. For example, “trump” will find “Trump” but also “Trump’s,” “Trumpism,” “Trumpian,” etc.

```
> #Cleanup
> rm(Thisfetch,
+     dates,
+     enddate
```

Week of

```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source

118
119 #####
120 #Headline coding
121 searchterms <- "trump|maga"
122 Headlines$Trump <- ifelse(grepl(searchterms,
123                               Headlines$title,
124                               ignore.case = TRUE), 1, 0)
125 searchterms <- "house|gop|republican|mccarthy"
126 Headlines$Republicans <- ifelse(grepl(searchterms,
127                                       Headlines$title,
128                                       ignore.case = TRUE), 1, 0)
129 searchterms <- "classified documents"
130 Headlines$Documents <- ifelse(grepl(searchterms,
131                                       Headlines$title,
132                                       ignore.case = TRUE), 1, 0)
133 searchterms <- "debt|ceiling"
134 Headlines$Debt <- ifelse(grepl(searchterms,
135
```

Replicate the code – with those two aspects altered – for each headline categorization you want to do.

This code creates a column called “Republicans” ...

... and codes it as a “1” if the headline includes any of these terms.

```
Console Terminal Background Jobs
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] 0 rows
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> CountsBySource <- Headlines %>%
+   group_by(Source) %>%
+   summarize(HeadlineCount = n())
> view(CountsBySource)
> filename <- paste0(query, startdate, "to", enddate, ".csv")
> write_excel_csv(Headlines, filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate
```



```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDEL scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source
187
188 #####
189 #Graphing by week and by day
190 if (!require("plotly")) install.packages("tidyverse")
191 library(plotly)
192 # color palettes: https://colors.co/palettes/trending
193 # graphing by week
194 fig <- plot_ly(AggByweek, x = ~weekof, y = ~Trump,
195               name = 'Trump', type = 'scatter',
196               mode = 'none', stackgroup = 'one',
197               fillcolor = '#F94144')
198 fig <- fig %>% add_trace(y = ~Republicans,
199                          name = 'Republicans',
200                          fillcolor = '#F3722C')
201 fig <- fig %>% add_trace(y = ~Documents, name = 'Documents',
202                          fillcolor = '#F8961E')
203 fig <- fig %>% add_trace(y = ~Debt, name = 'Debt',
204                          fillcolor = '#F94144')
```

Environment History Connections Tutorial

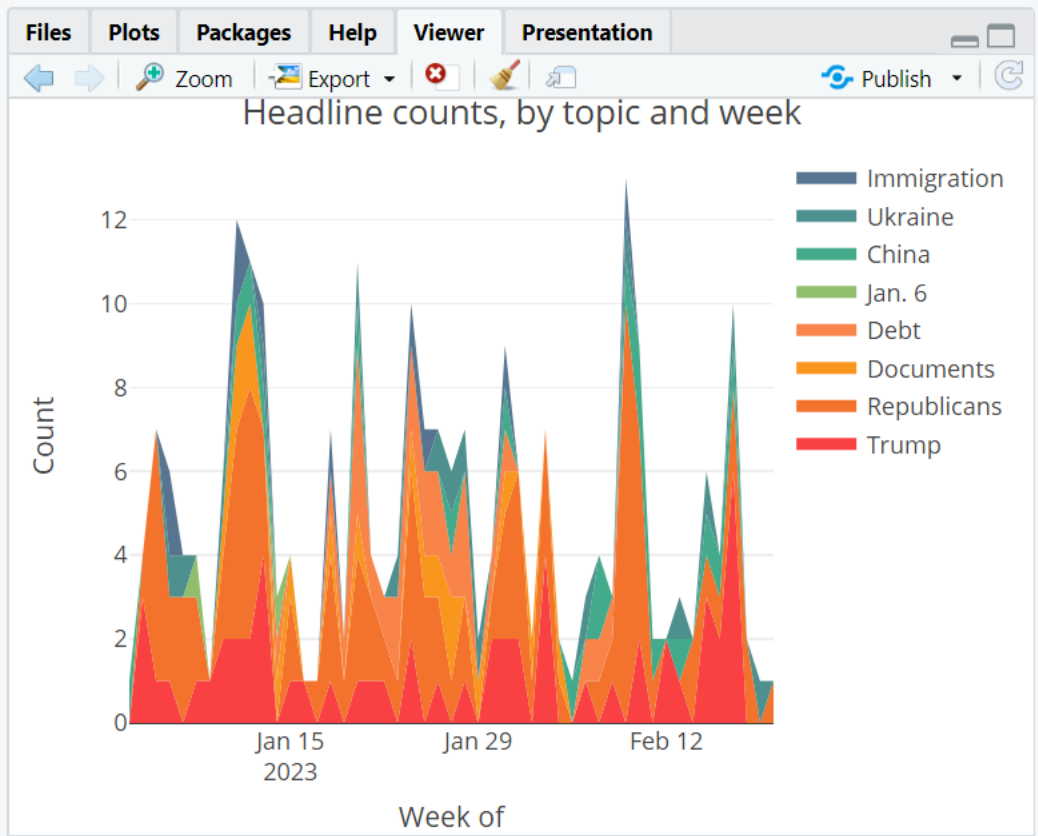
R Global Environment 422 MiB

Data

AggByDay	49 obs. of 10 variables
AggByweek	8 obs. of 10 variables
CountsBySour...	2 obs. of 2 variables
fig	List of 8
Headlines	543 obs. of 15 variables
wordCounts	1726 obs. of 2 variables

Values

Hours	0.1
-------	-----



Also, be sure each headline categorization column's name gets added to the chart code, along with ...

```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDELt scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source
187
188 #####
189 #Graphing by week and by day
190 if (!require("plotly")) install.packages("tidyverse")
191 library(plotly)
192 # color palettes: https://colors.co/palettes/trending
193 # Graphing by week
194 fig <- plot_ly(AggByweek, x = ~weekof, y = ~Trump,
195               name = 'Trump', type = 'scatter',
196               mode = 'none', stackgroup = 'one',
197               fillcolor = '#F94144')
198 fig <- fig %>% add_trace(y = ~Republicans,
199                       name = 'Republicans',
200                       fillcolor = '#F3722C')
201 fig <- fig %>% add_trace(y = ~Documents, name = 'Documents',
202                       fillcolor = '#F8961E')
203 fig <- fig %>% add_trace(y = ~Debt, name = 'Debt',
204                       fillcolor = '#F94144')
```

Environment History Connections Tutorial

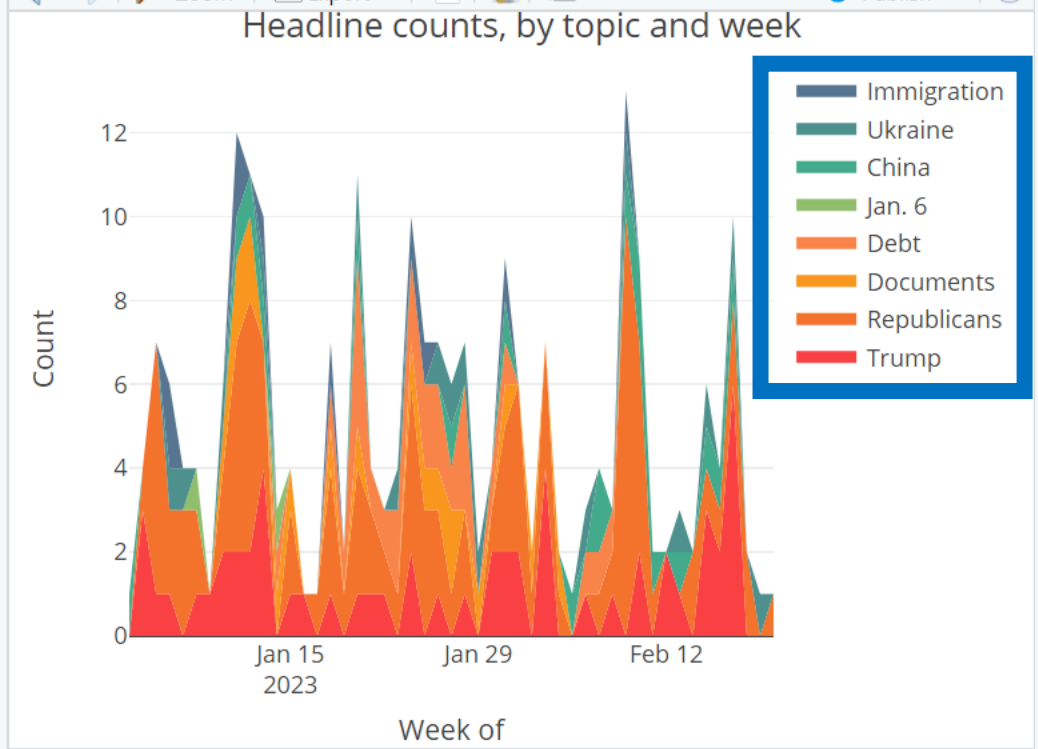
R Global Environment 422 MiB

Data

AggByDay	49 obs. of 10 variables
AggByweek	8 obs. of 10 variables
CountsBySour...	2 obs. of 2 variables
fig	List of 8

... labels for the chart's legend.

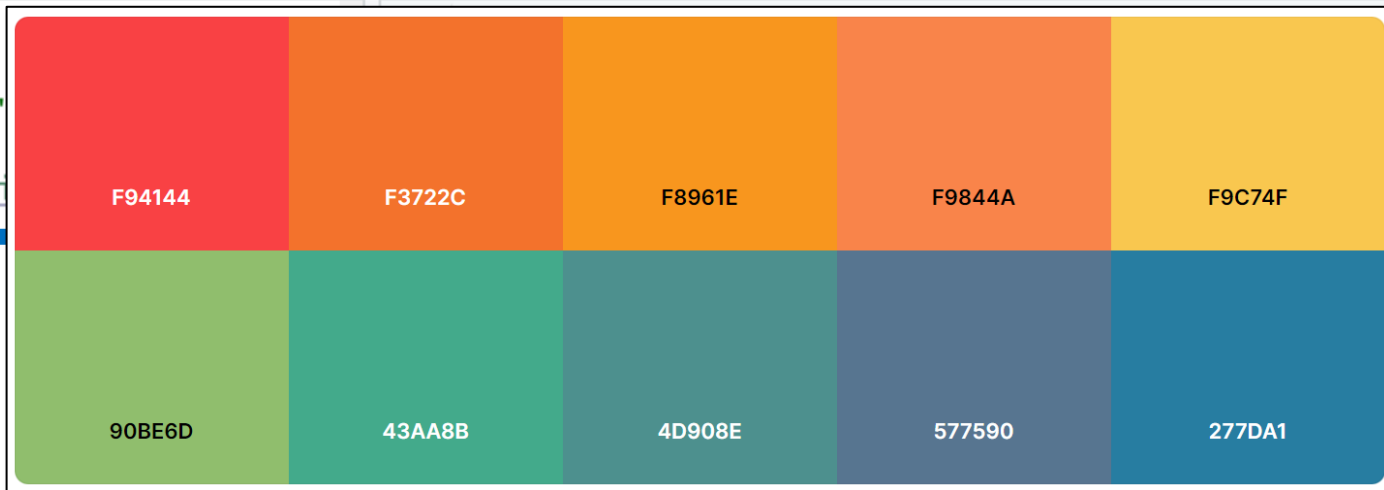
Also, be sure each headline categorization column's name gets added to the chart code, along with ...



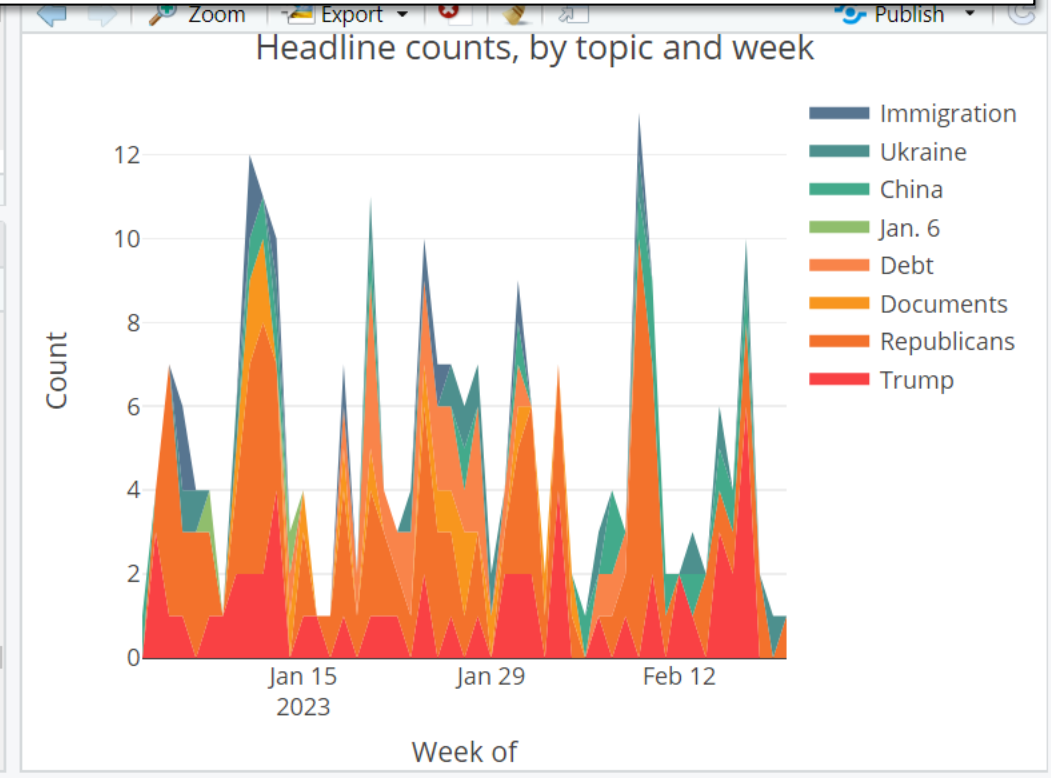
```

187
188 #####
189 #Graphing by week and by day
190 if (!require("plotly")) install.packages("tidyverse")
191 library(plotly)
192 # color palettes: https://colors.co/palettes/trend
193 # Graphing by week
194 fig <- plot_ly(AggByweek, x = ~weekof, y = ~Trump,
195               name = 'Trump', type = 'scatter',
196               mode = 'none', stackgroup = 'one',
197               fillcolor = '#F94144')
198 fig <- fig %>% add_trace(y = ~Republicans,
199                       name = 'Republicans',
200                       fillcolor = '#F3722C')
201 fig <- fig %>% add_trace(y = ~Documents, name = 'Documents',
202                       fillcolor = '#F8961E')
203 fig <- fig %>% add_trace(y = ~Debt, name = 'Debt',
204                       fillcolor = '#F9C44A')

```



Edit these color codes to change the chart's colors. I used a palette from Colors.co.



```

> filename <- paste0(query, startdate, to, enddate, '.csv')
> write_excel_csv(Headlines, filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate

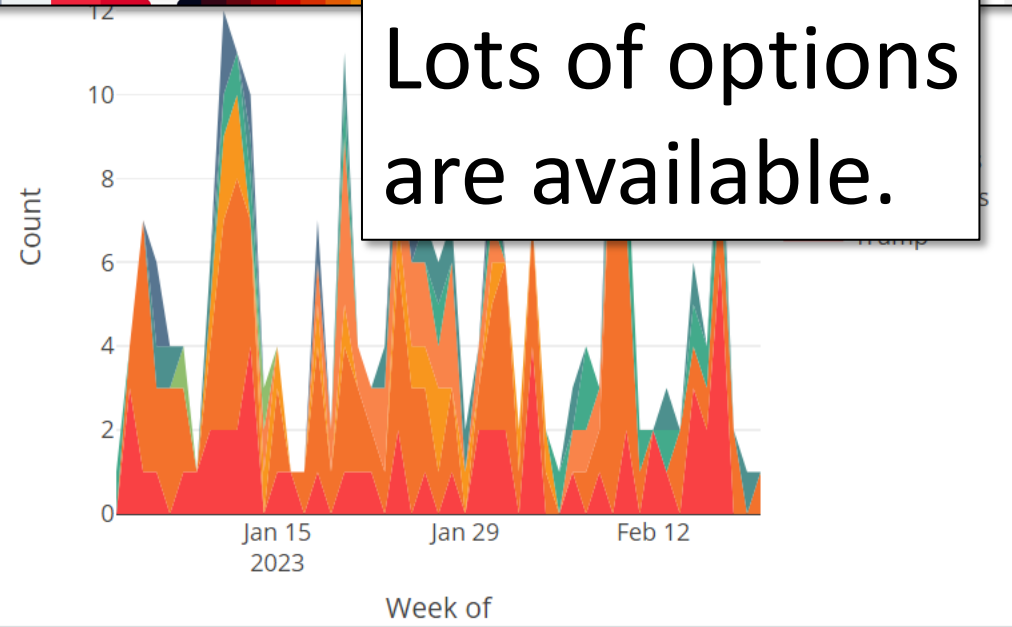
```

```
GDELT scraper.R x CountsBySource x Headlines x WordCounts x  
Source on Save Run  
187  
188 #####  
189 #Graphing by week and by day  
190 if (!require("plotly")) install.packages("tidyvers  
191 library(plotly)  
192 # color palettes: https://colors.co/palettes/tren  
193 # Graphing by week  
194 fig <- plot_ly(AggByweek, x = ~weekof, y = ~Trump,  
195 name = 'Trump', type = 'scatter',  
196 mode = 'none', stackgroup = 'one',  
197 fillcolor = '#F94144')  
198 fig <- fig %>% add_trace(y = ~Republicans,  
199 name = 'Republicans',  
200 fillcolor = '#F3722C')  
201 fig <- fig %>% add_trace(y = ~Documents, name = 'D  
202 fillcolor = '#F8961E')  
203 fig <- fig %>% add_trace(y = ~Debt, name = 'Debt',  
204 fillcolor = '#504441')
```

The screenshot shows the Coolors website interface. At the top, there are navigation tabs for 'Environment', 'History', 'Connections', and 'Tutorial'. Below the navigation is a search bar with the text 'Search with colors, topics, styles or hex values...'. The main heading is 'Trending Color Palettes' with a subtext 'Get inspired by thousands of beautiful color schemes and make something cool!'. The page displays a grid of various color palettes, each with a set of colored squares and a heart icon indicating popularity. Some palettes include a 'Nucleo' logo, which is a library of 29280 icons.

Edit these color codes to change the chart's colors. I used a palette from Colors.co.

Lots of options are available.



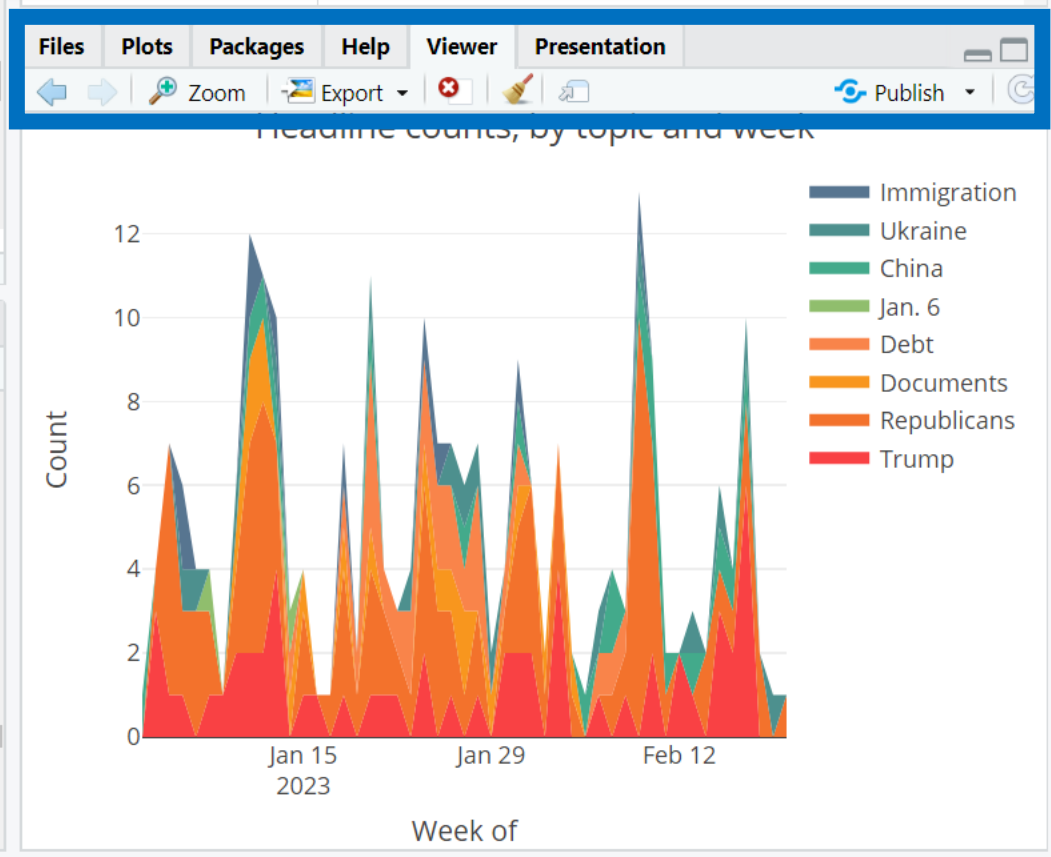
```
> filename <- paste0(query, startdate, to, enddate, '.csv')  
> write_excel_csv(Headlines, filename)  
> #Cleanup  
> rm(Thisfetch,  
+ dates,  
+ enddate
```

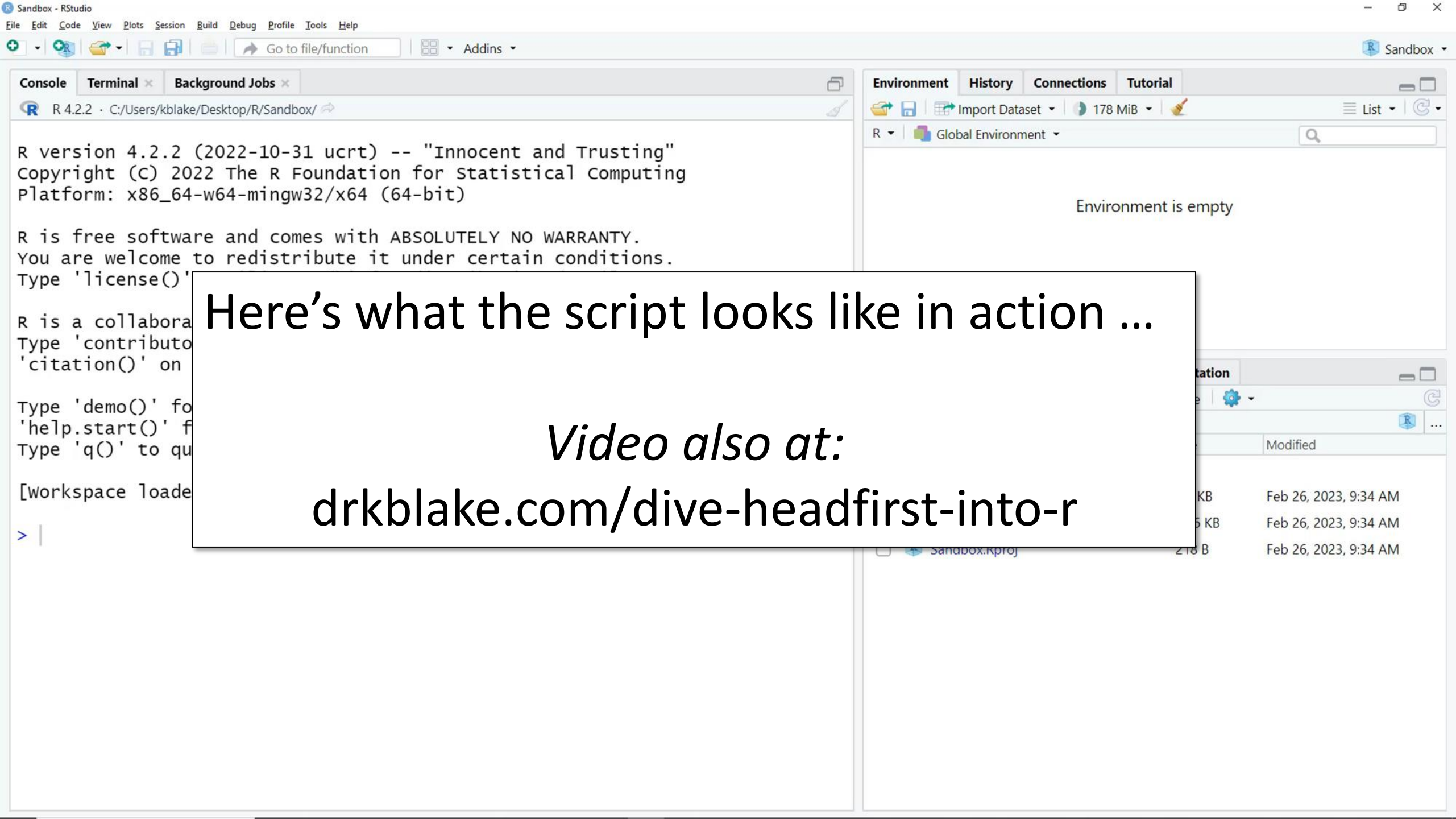
```
Sandbox - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

GDELT scraper.R x CountsBySource x Headlines x WordCounts x
Source on Save Run Source
187
188 #####
189 #Graphing by week and by day
190 if (!require("plotly")) install.packages("tidyverse")
191 library(plotly)
192 # color palettes: https://colors.co/palettes/trending
193 # Graphing by week
194 fig <- plot_ly(AggByweek, x = ~weekof, y = ~Trump,
195               name = 'Trump', type = 'scatter',
196               mode = 'none', stackgroup = 'one',
197               fillcolor = '#F94144')
198 fig <- fig %>% add_trace(y = ~Republicans,
199                          name = 'Republicans',
200                          fillcolor = '#F3722C')
201 fig <- fig %>% add_trace(y = ~Documents, name = 'Documents',
202                          fillcolor = '#F8961E')
203 fig <- fig %>% add_trace(y = ~Debt, name = 'Debt',
204                          fillcolor = '#F08080')
```

```
Console Terminal Background Jobs
R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/
[1] U R OWS
> #Deduplicate, check, and export data
> Headlines <- Headlines[!duplicated(Headlines$URL),]
> CountsBySource <- Headlines %>%
+   group_by(source) %>%
+   summarize(Headlinecount = n())
> view(CountsBySource)
> filename <- paste0(query,startdate,"to",enddate,".csv")
> write_excel_csv(Headlines,filename)
> #Cleanup
> rm(Thisfetch,
+   dates,
+   enddate
```

Once made, the chart can be copied, exported as .html, or published (for free) on RPubS.





Here's what the script looks like in action ...

Video also at:

drkblake.com/dive-headfirst-into-r

Console Terminal Background Jobs

R 4.2.2 · C:/Users/kblake/Desktop/R/Sandbox/

```
R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"  
Copyright (C) 2022 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

```
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/kblake/Desktop/R/Sandbox/.RData]
```

```
> |
```

Environment History Connections Tutorial

Import Dataset 178 MiB

R Global Environment

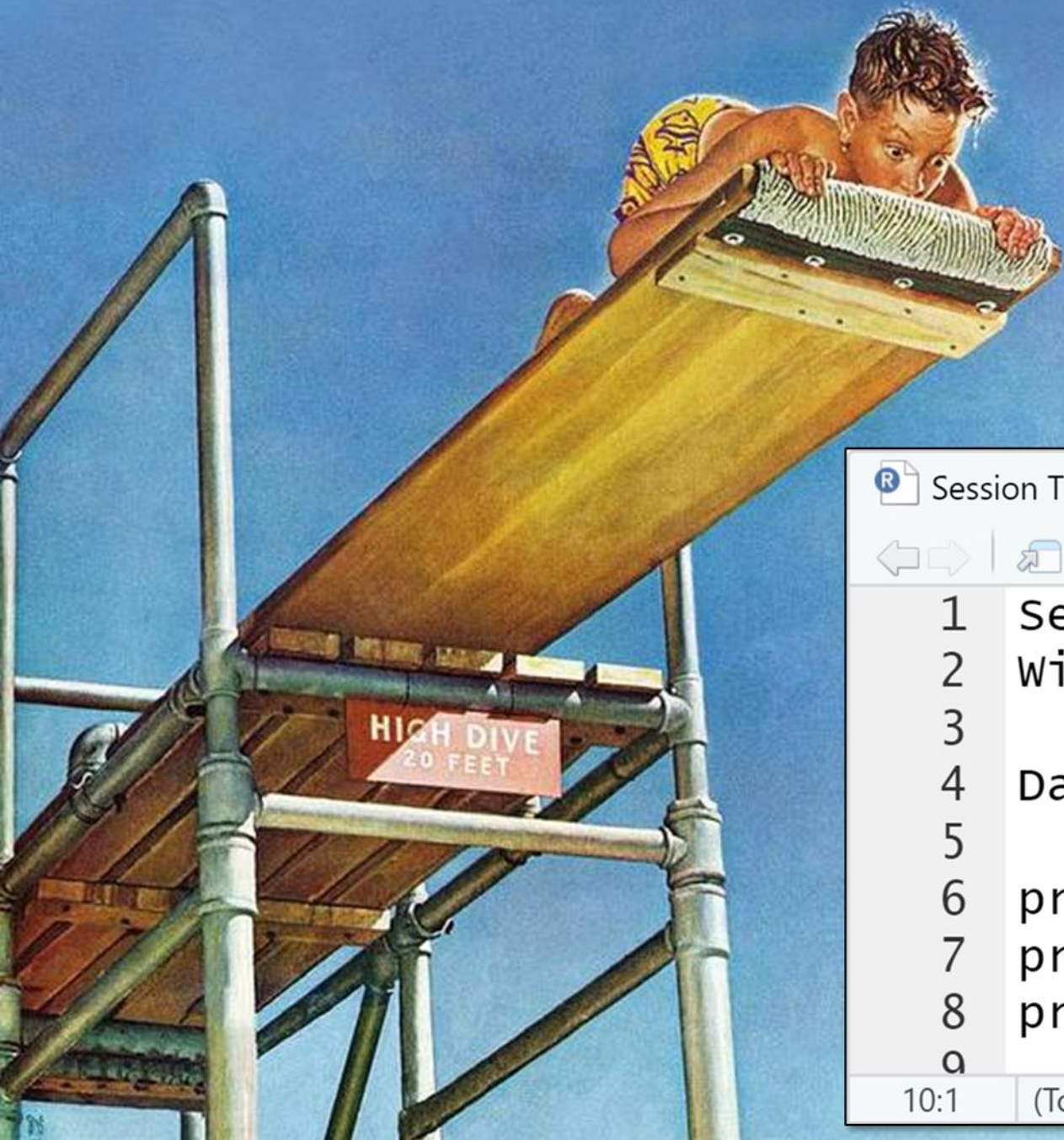
Environment is empty

Files Plots Packages Help Viewer Presentation

Folder Blank File Delete Rename

C: > Users > kblake > Desktop > R > Sandbox

	Name	Size	Modified
	..		
<input type="checkbox"/>	.RData	2.5 KB	Feb 26, 2023, 9:34 AM
<input type="checkbox"/>	.Rhistory	14.6 KB	Feb 26, 2023, 9:34 AM
<input type="checkbox"/>	Sandbox.Rproj	218 B	Feb 26, 2023, 9:34 AM



```
Session Title.R x
Source on Save
1 Session <- "Dive Headfirst into R"
2 With <- c("Dr. Jun Zhang",
3           "Dr. Ken Blake")
4 Date <- as.Date("03/02/2023",
5                 "%m/%d/%y")
6 print(Session)
7 print(With)
8 print(Date)
9
10:1 (Top Level) R Script
```