R and tidyverse

PEETER TINITS

#DIGMET SUMMER SCHOOL, TARTU

28.08.2019

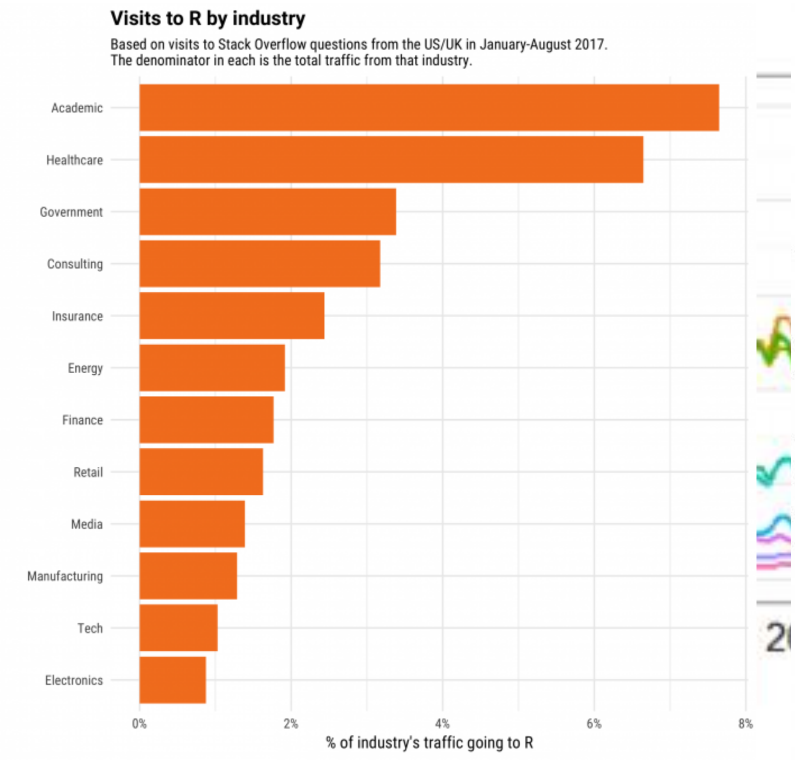
R language

R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing.

The R language is widely used among statisticians and data miners for developing statistical software[7] and data analysis.[8] Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.[9] R ranks 8th in the TIOBE index.

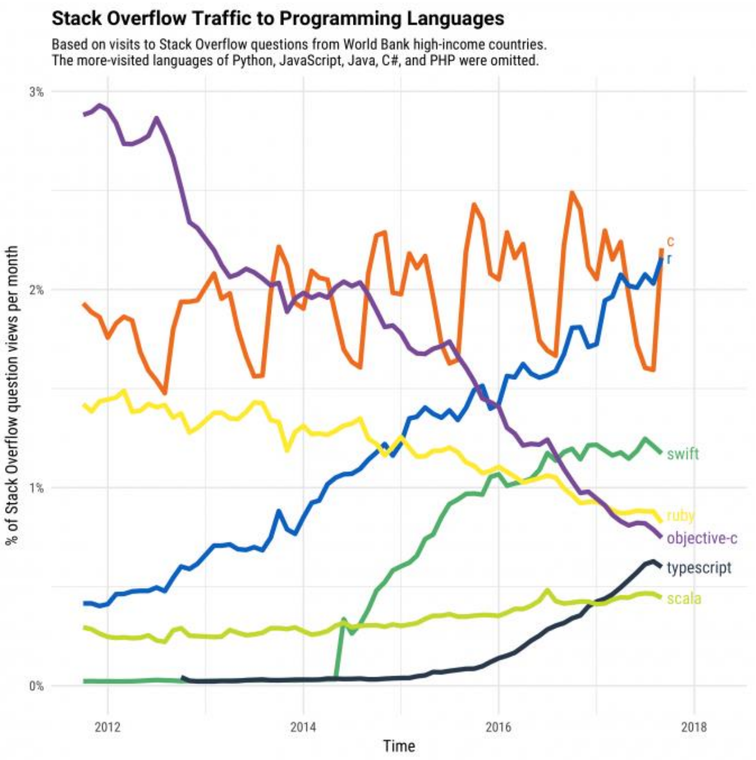
-Wikipedia

A growing community



Visits to R by industry

Based on visits to Stack Overflow questions from the US/UK in January-August 2017.
The denominator in each is the total traffic from that industry.



Stack Overflow Traffic to Programming Languages

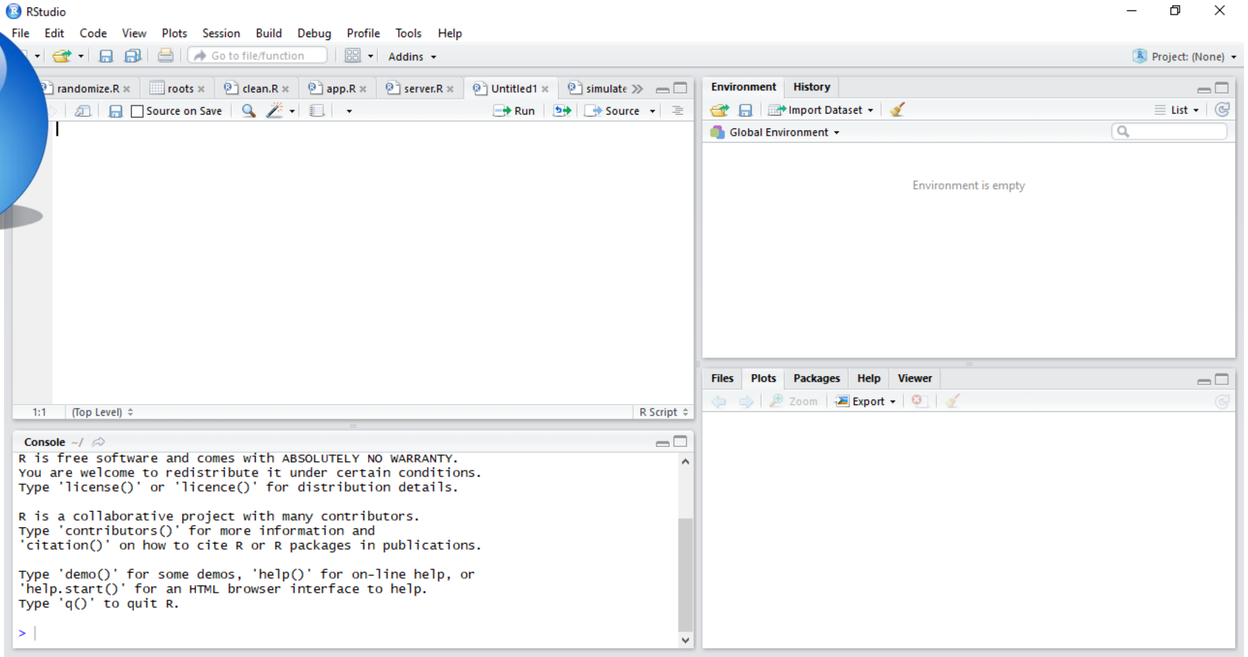
Based on visits to Stack Overflow questions from World Bank high-income countries.
The more-visited languages of Python, JavaScript, Java, C#, and PHP were omitted.

https://stackoverflow.blog/2017/10/10/impressive-growth-r/

RStudio

Rstudio – free and open source integrated developing environment.

Your window to R!



```
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.

> 
```

Tidyverse

Tidyverse:

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

- https://www.tidyverse.org/

> Programs must be written for people to read, and only incidentally for machines to execute.
> — Hal Abelson

Working with data

When working with data you must:

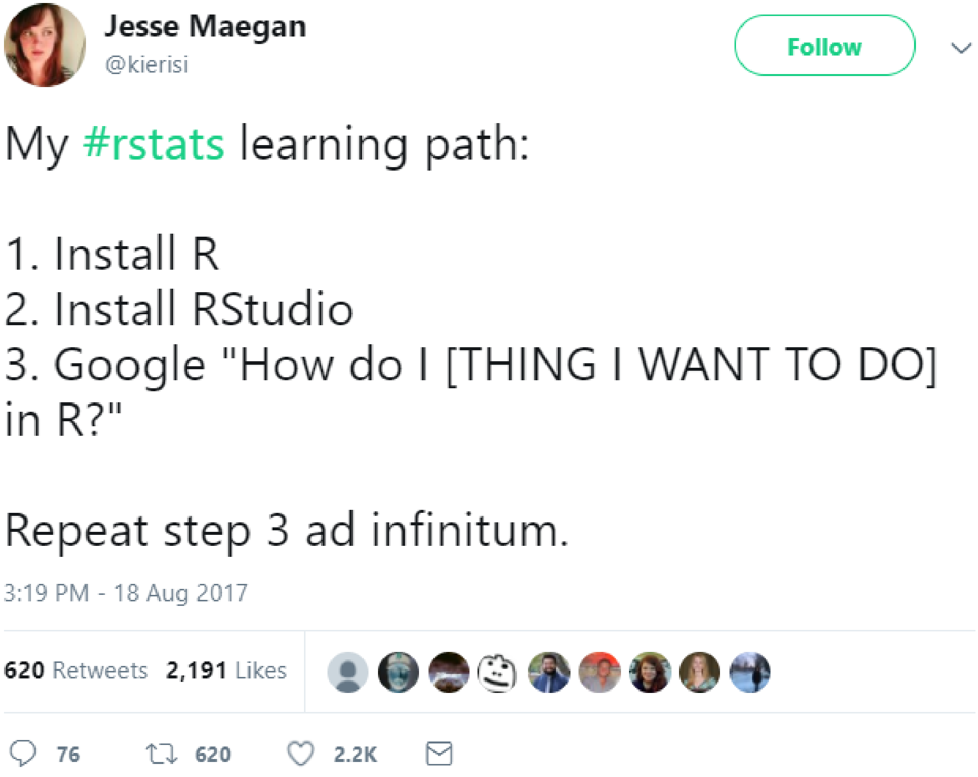
- Figure out what you want to do.
- Describe those tasks in the form of a computer program.
- Execute the program.

Tidyverse makes these steps fast and easy:

- Offers simple and intuitive options
- Data manipulation is organized with verbs.
- Quicker than traditional R.

https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html

How to use R



Jesse Maegan
@kierisi

My #rstats learning path:

1. Install R
2. Install RStudio
3. Google "How do I [THING I WANT TO DO] in R?"

Repeat step 3 ad infinitum.

3:19 PM - 18 Aug 2017

620 Retweets 2,191 Likes

Let’s get started

Get the data

Main page:

http://tiny.cc/digmetTM (or click here)

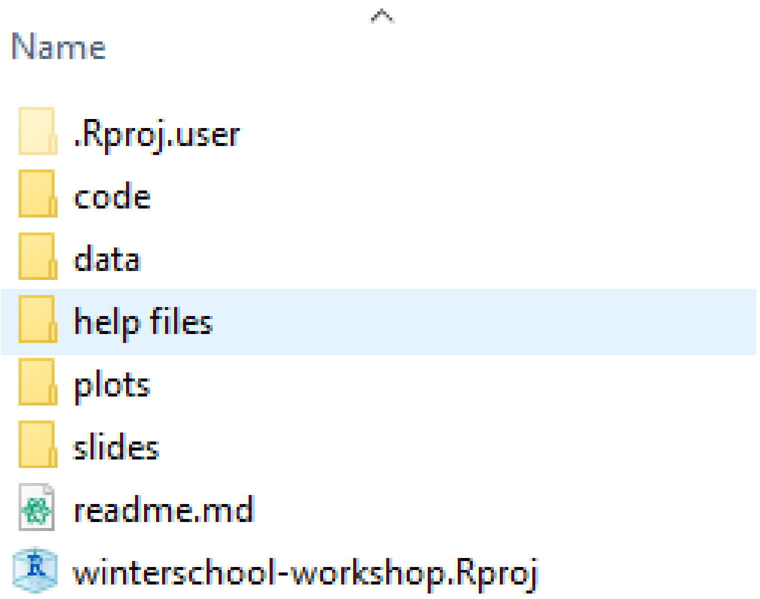
To get the files:

1. Click on the download .zip link in the top right green button, or click here.
2. Unpack the files where you want them.

Alternatively use Git (instructions on the site).

Project files

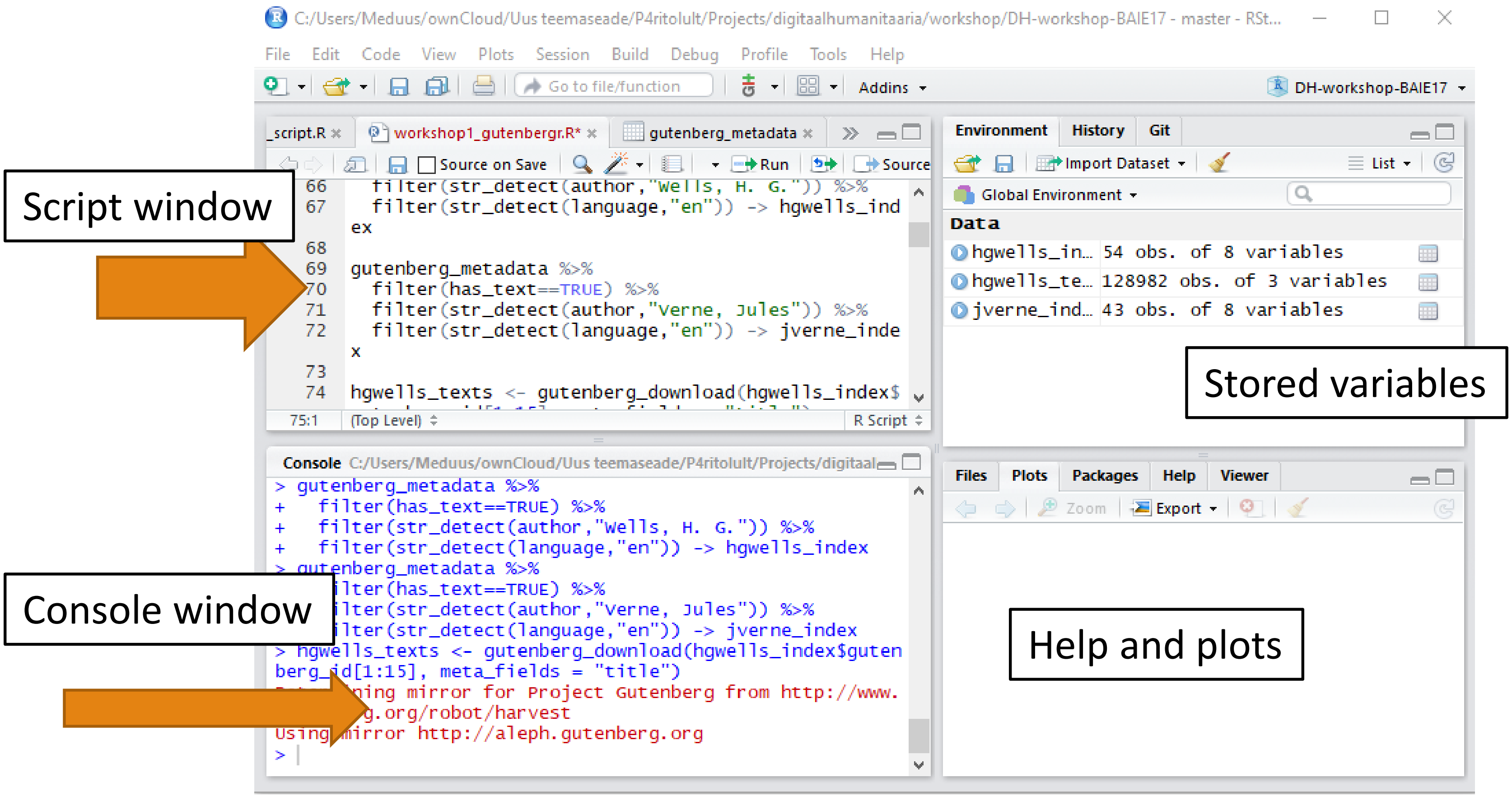
To get started, run the Rproj file. This simplif things in Rstudio, sets the working directory and remembers your actions.



Name

- .Rproj.user
- code
- data
- help files
- plots
- slides
- readme.md
- winterschool-workshop.Rproj

RStudio view



C:/Users/Meduus/ownCloud/Uus teemaseade/P4ritolult/Projects/digitaalhumanitaaria/workshop/DH-workshop-BAIE17 - master - RSt...
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
DH-workshop-BAIE17
_script.R
workshop1_gutenbergr.R*
gutenberg_metadata
Source on Save
Run
Source
75:1
(Top Level)
R Script
Console C:/Users/Meduus/ownCloud/Uus teemaseade/P4ritolult/Projects/digitaal
Environment
History
Git
Import Dataset
List
Global Environment
Files
Plots
Packages
Help
Viewer
Zoom
Export

```
  filter(str_detect(author,"Wells, H. G.")) %>%
  filter(str_detect(language,"en")) -> hgwells_index

gutenberg_metadata %>%
  filter(has_text==TRUE) %>%
  filter(str_detect(author,"Verne, Jules")) %>%
  filter(str_detect(language,"en")) -> jverne_index

hgwells_texts <- gutenberg_download(hgwells_index$
```

Data

```
hgwells_in…  54 obs. of 8 variables
hgwells_te…  128982 obs. of 3 variables
jverne_ind…  43 obs. of 8 variables
```

```
> gutenberg_metadata %>%
+   filter(has_text==TRUE) %>%
+   filter(str_detect(author,"Wells, H. G.")) %>%
+   filter(str_detect(language,"en")) -> hgwells_index
> gutenberg_metadata %>%
+   filter(has_text==TRUE) %>%
+   filter(str_detect(author,"Verne, Jules")) %>%
+   filter(str_detect(language,"en")) -> jverne_index
> hgwells_texts <- gutenberg_download(hgwells_index$gutenberg_id[1:15], meta_fields = "title")
Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
Using mirror http://aleph.gutenberg.org
> 
```

Script window

Console window

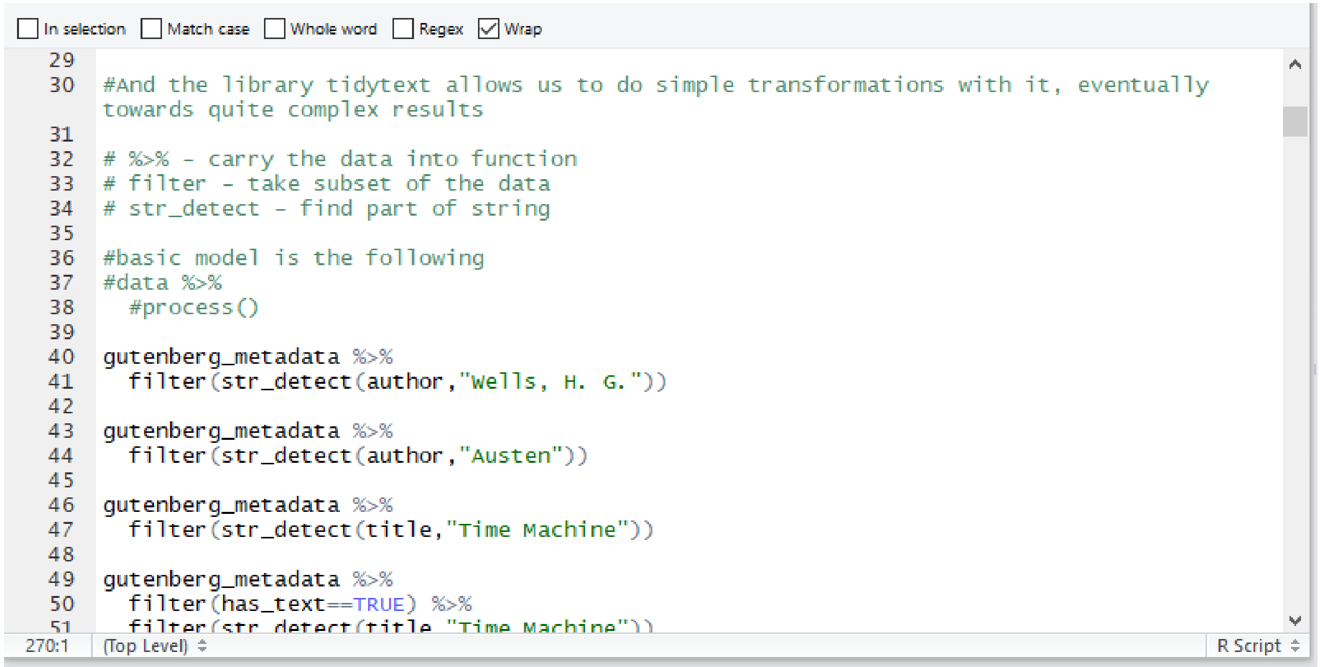
Stored variables

Help and plots

Script files

Green = comments, (and text strings – e.g. „Wells, H. G.“)

black,blue, etc = code



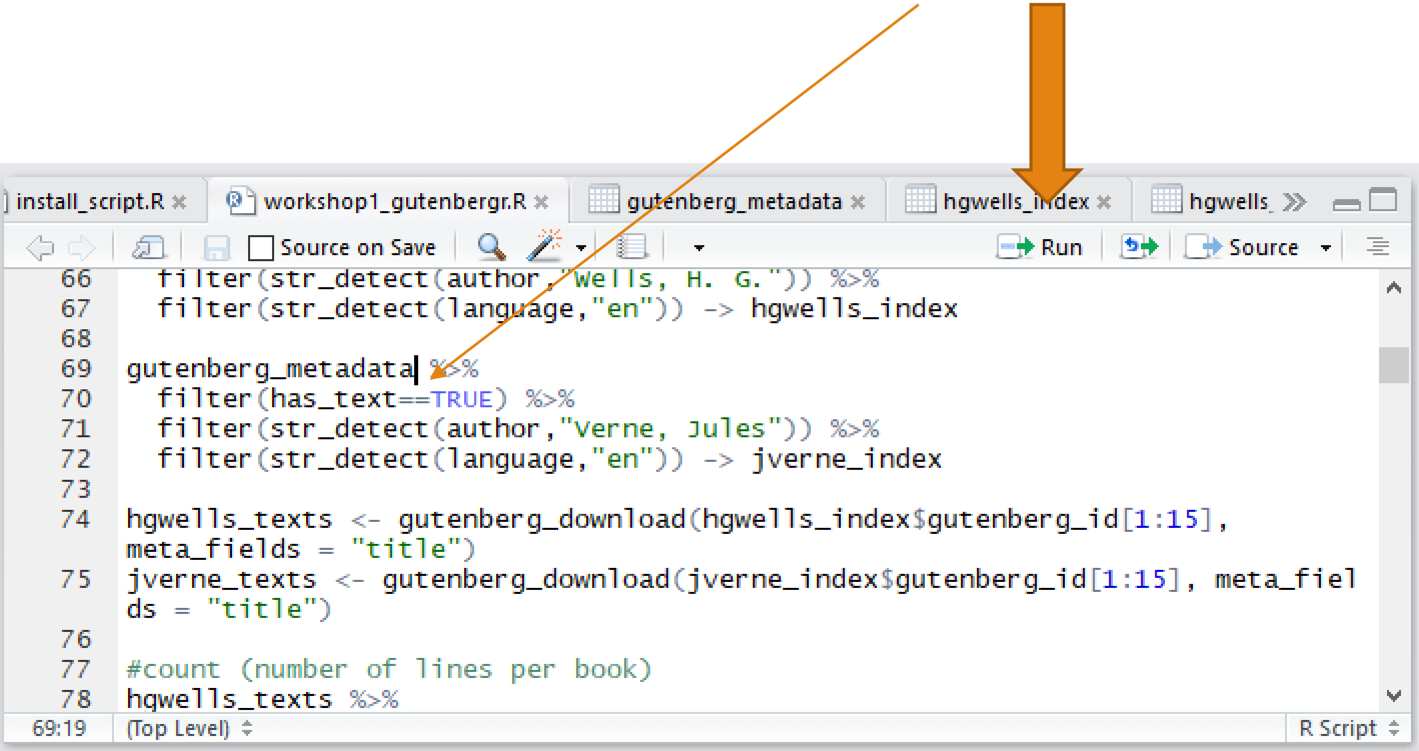
□ In selection □ Match case □ Whole word □ Regex ☑ Wrap

```

#And the library tidytext allows us to do simple transformations with it, eventually towards quite complex results

# %>% - carry the data into function
# filter - take subset of the data
# str_detect - find part of string

#basic model is the following
#data %>%
  #process()

gutenberg_metadata %>%
  filter(str_detect(author,"Wells, H. G."))

gutenberg_metadata %>%
  filter(str_detect(author,"Austen"))

gutenberg_metadata %>%
  filter(str_detect(title,"Time Machine"))

gutenberg_metadata %>%
  filter(has_text==TRUE) %>%
  filter(str_detect(title,"Time Machine"))
```

Basic R code

To run code

Pick a line and click run

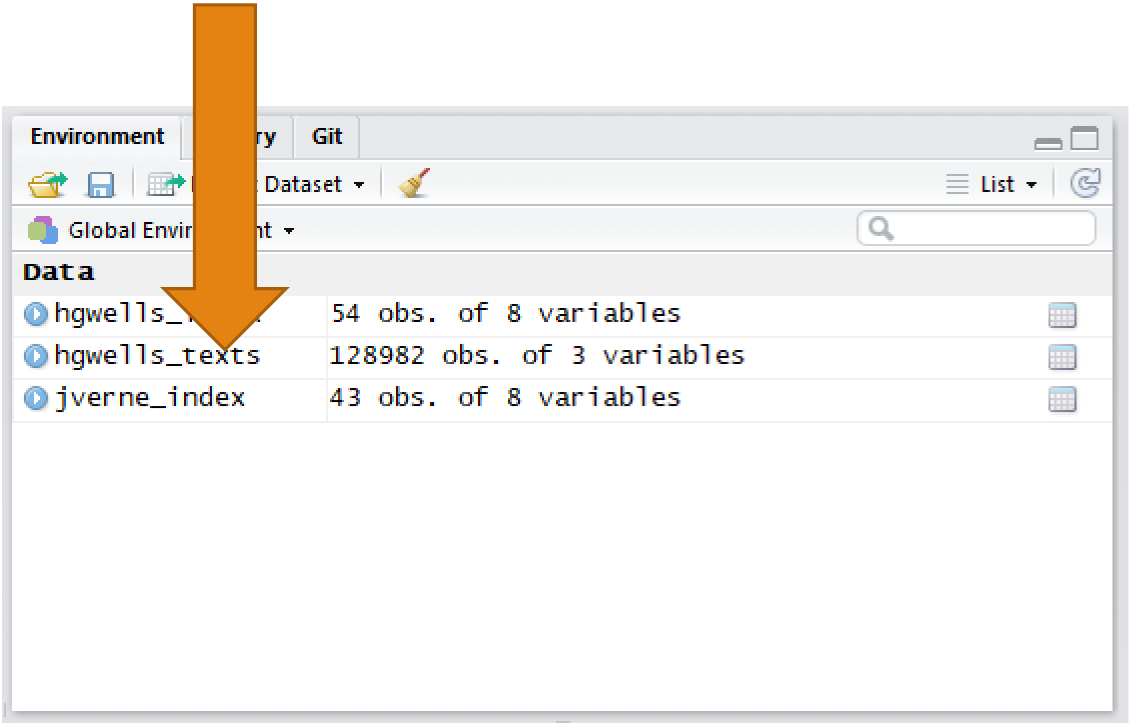


```
  filter(str_detect(author,"Wells, H. G.")) %>%
  filter(str_detect(language,"en")) -> hgwells_index

gutenberg_metadata %>%
  filter(has_text==TRUE) %>%
  filter(str_detect(author,"Verne, Jules")) %>%
  filter(str_detect(language,"en")) -> jverne_index

hgwells_texts <- gutenberg_download(hgwells_index$gutenberg_id[1:15], meta_fields = "title")
jverne_texts <- gutenberg_download(jverne_index$gutenberg_id[1:15], meta_fields = "title")

#count (number of lines per book)
hgwells_texts %>%
```

Data

To look at the data

To view a dataframe

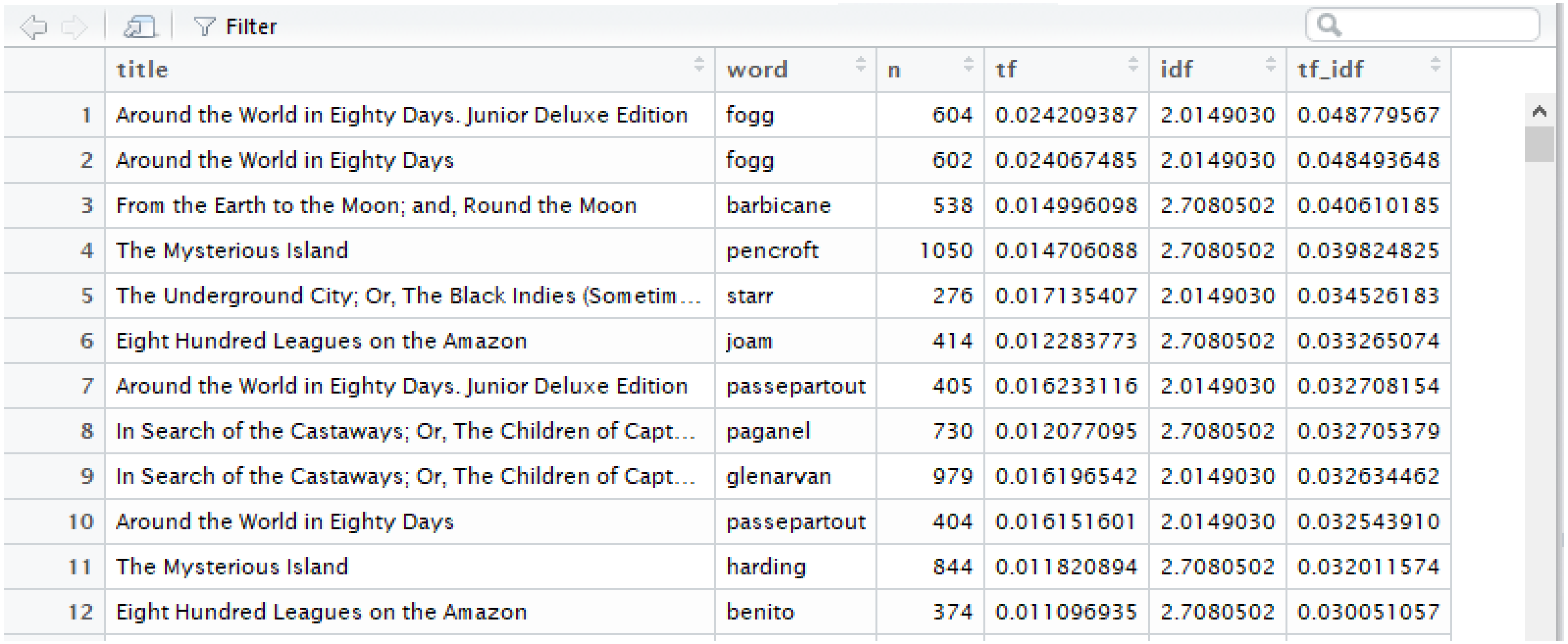


Data

```
hgwells_           54 obs. of 8 variables
hgwells_texts      128982 obs. of 3 variables
jverne_index       43 obs. of 8 variables
```

How data looks like

Just a table really ☺

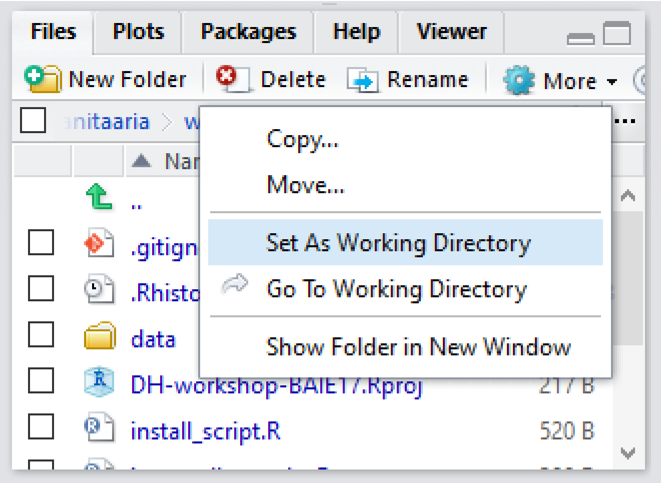


	title	word	n	tf	idf	tf_idf
1	Around the World in Eighty Days. Junior Deluxe Edition	fogg	604	0.024209387	2.0149030	0.048779567
2	Around the World in Eighty Days	fogg	602	0.024067485	2.0149030	0.048493648
3	From the Earth to the Moon; and, Round the Moon	barbicane	538	0.014996098	2.7080502	0.040610185
4	The Mysterious Island	pencroft	1050	0.014706088	2.7080502	0.039824825
5	The Underground City; Or, The Black Indies (Sometim…	starr	276	0.017135407	2.0149030	0.034526183
6	Eight Hundred Leagues on the Amazon	joam	414	0.012283773	2.7080502	0.033265074
7	Around the World in Eighty Days. Junior Deluxe Edition	passepartout	405	0.016233116	2.0149030	0.032708154
8	In Search of the Castaways; Or, The Children of Capt…	paganel	730	0.012077095	2.7080502	0.032705379
9	In Search of the Castaways; Or, The Children of Capt…	glenarvan	979	0.016196542	2.0149030	0.032634462
10	Around the World in Eighty Days	passepartout	404	0.016151601	2.0149030	0.032543910
11	The Mysterious Island	harding	844	0.011820894	2.7080502	0.032011574
12	Eight Hundred Leagues on the Amazon	benito	374	0.011096935	2.7080502	0.030051057

To read files

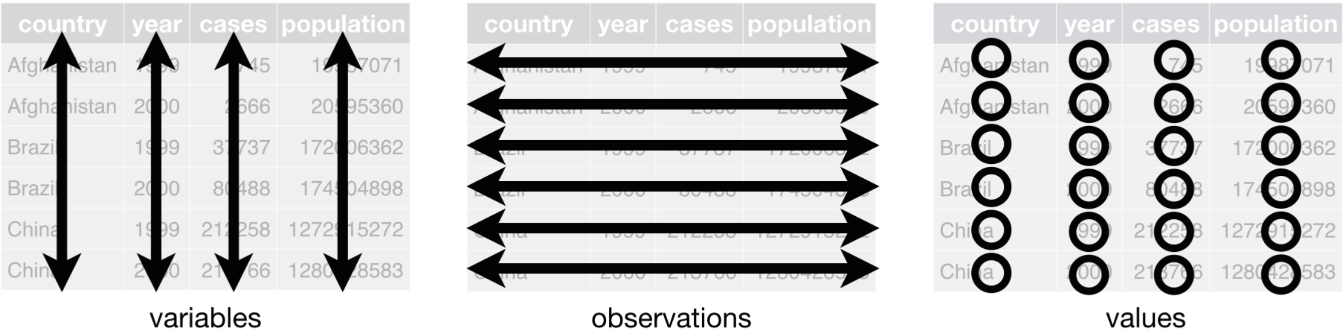
If you do not use Rproject:

- You need to set the right working directory to find data



Tidy data

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.



variables

observations

values

http://vita.had.co.nz/papers/tidy-data.html

Basic tidyverse operations

%>% - pushes the result to be processed on the next line

```
data %>%
    process()
```

select() - selecting variables

filter() - provides basic filtering capabilities

group_by() - groups data by categorical levels

summarise() - summarise data by functions of choice

arrange() - ordering data

join() - joining separate dataframes

mutate() - create new variables

Tidyverse sources

Wickham, H. 2017. Tidy tools Manifesto

Wichkah, H. In progress. The Tidyverse Style Guide

Korde, R. 2017. Switching from base R to tidyverse

Boehmke, B. 2015. Data Processing with dplyr & tidyr

Grolemund, G.; Wickham, H. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data

Silge, J.; Robinson, D. 2017. Text Mining with R. A tidy approach.

Extra: Tidyverse alternative 1

Nested Option:

```
arrange(
        summarize(
            filter(data, variable == numeric_value),
            Total = sum(variable)
        ),
    desc(Total)
)
```

https://rpubs.com/bradleyboehmke/data_wrangling

Extra: Tidyverse alternative 2

Multiple Object Option:

```
a <- filter(data, variable == numeric_value)
b <- summarise(a, Total = sum(variable))
c <- arrange(b, desc(Total))
```

https://rpubs.com/bradleyboehmke/data_wrangling

Extra: Tidyverse solution

%>% Option:

```
  data %>%
      filter(variable == “value”) %>%
      summarise(Total = sum(variable)) %>%
      arrange(desc(Total))
```

https://rpubs.com/bradleyboehmke/data_wrangling