

R and tidyverse

SIMPLE TEXT ANALYTICS WITH R AND TIDYTEXT

PEETER TINITS 07.02.2018

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.

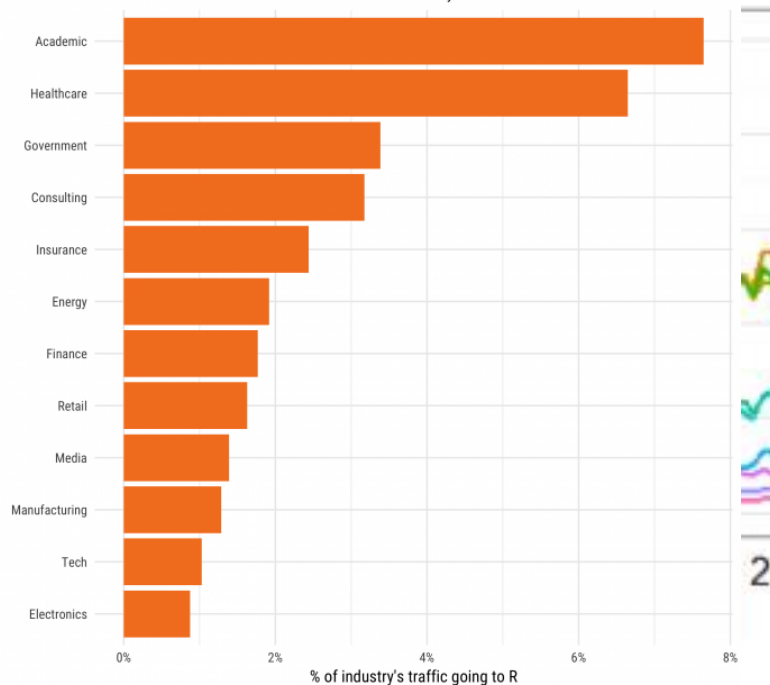
The R language is widely used among [statisticians](#) and [data miners](#) for developing [statistical software](#)^[7] and [data analysis](#).^[8] Polls, [surveys of data miners](#), and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.^[9]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

-Wikipedia

A growing community

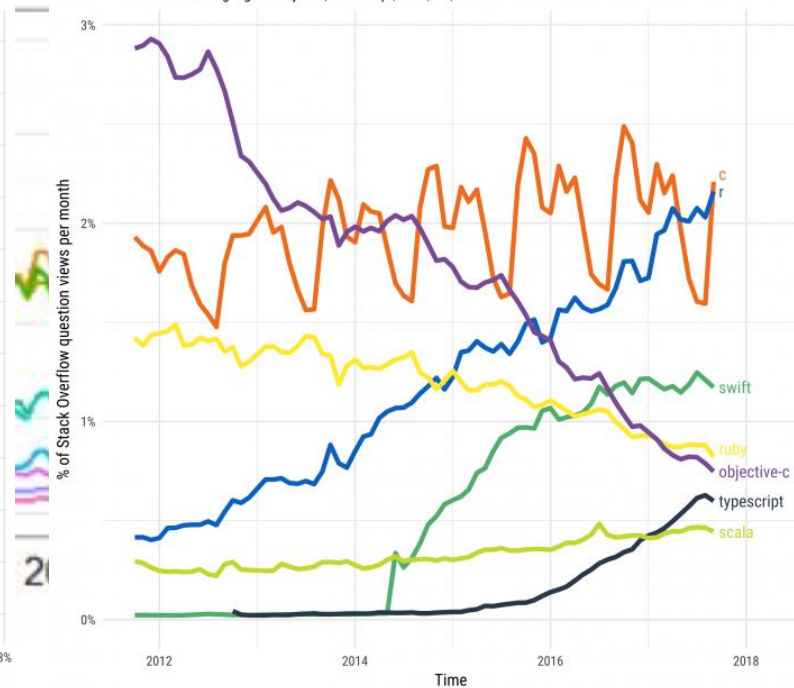
Visits to R by industry

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.

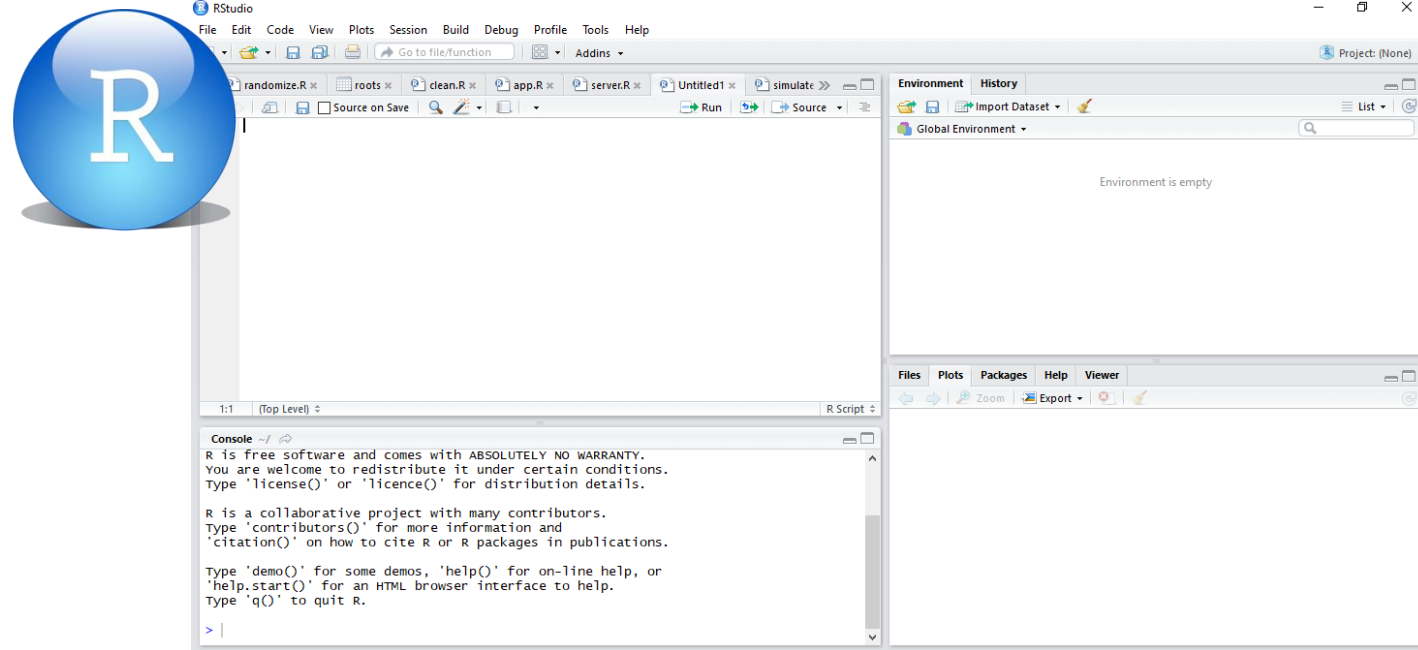


<https://stackoverflow.blog/2017/10/10/impressive-growth-r/>

RStudio

Rstudio – free and open source integrated developing environment.

Your window to R!



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

Follow



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



76

620

2.2K



Let's get started

Get the data

Main page:

<https://github.com/peeter-t2/DH-workshop-BAIE17>

1. Navigate to the folder where you want to keep the files, e.g. the downloads folder.
2. Run git
3. type "git clone <https://github.com/peeter-t2/DH-workshop-BAIE17.git>"

Alternatively:









1. Click on the download .zip link in the top right green button, or follow this link: <https://github.com/peeter-t2/DH-workshop-BAIE17/archive/master.zip>
2. Unpack the files where you want them.

To get started

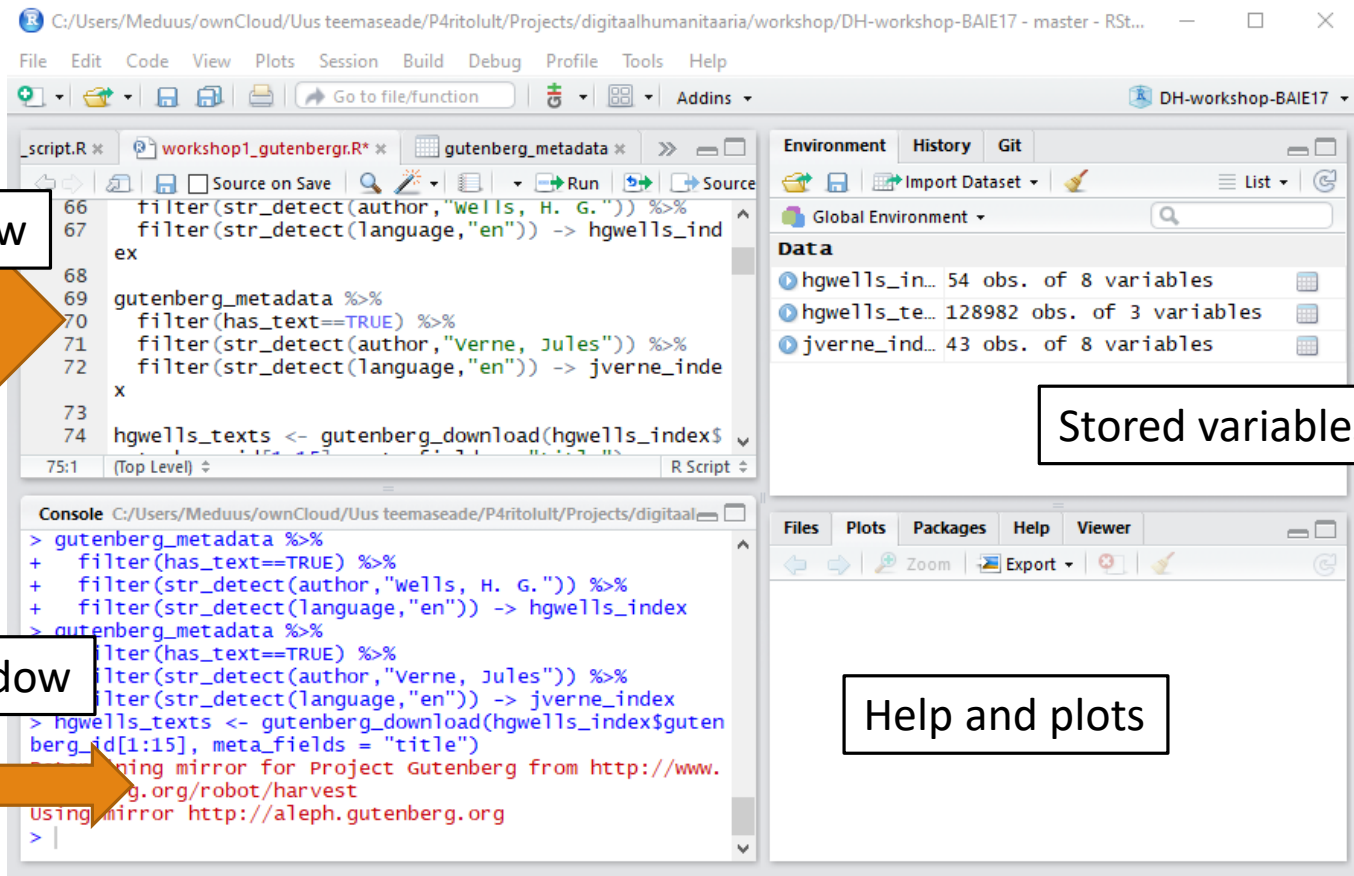
To get started, run the Rproj file. This simplifies things in Rstudio, sets the working directory and remembers your actions.



Name

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

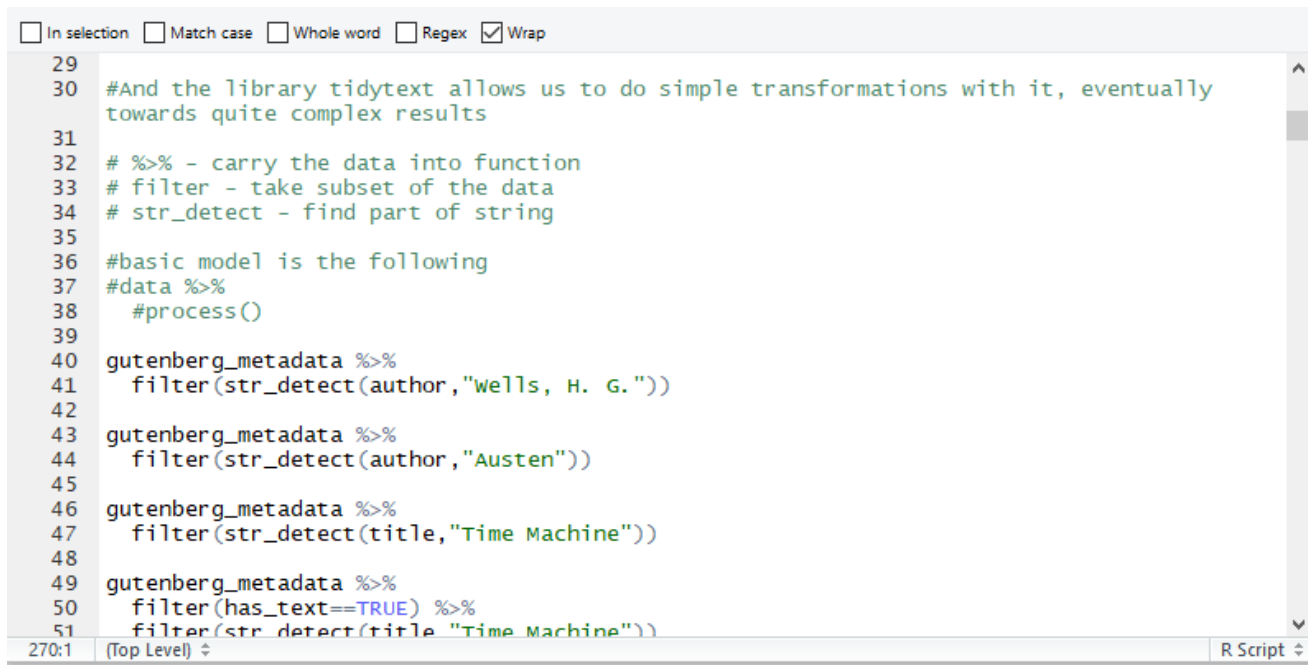
RStudio view



Script files

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

black,blue, etc = code



```
☐ In selection ☐ Match case ☐ Whole word ☐ Regex ☒ Wrap
29
30 #And the library tidytext allows us to do simple transformations with it, eventually
   towards quite complex results
31
32 # %>% - carry the data into function
33 # filter - take subset of the data
34 # str_detect - find part of string
35
36 #basic model is the following
37 #data %>%
38   #process()
39
40 gutenber_metadata %>%
41   filter(str_detect(author,"wells, H. G. "))
42
43 gutenber_metadata %>%
44   filter(str_detect(author,"Austen"))
45
46 gutenber_metadata %>%
47   filter(str_detect(title,"Time Machine"))
48
49 gutenber_metadata %>%
50   filter(has_text==TRUE) %>%
51   filter(str_detect(title,"Time Machine"))
```

270:1 (Top Level) ↕ R Script ↕

Basic R code

To run code

Pick a line and click run



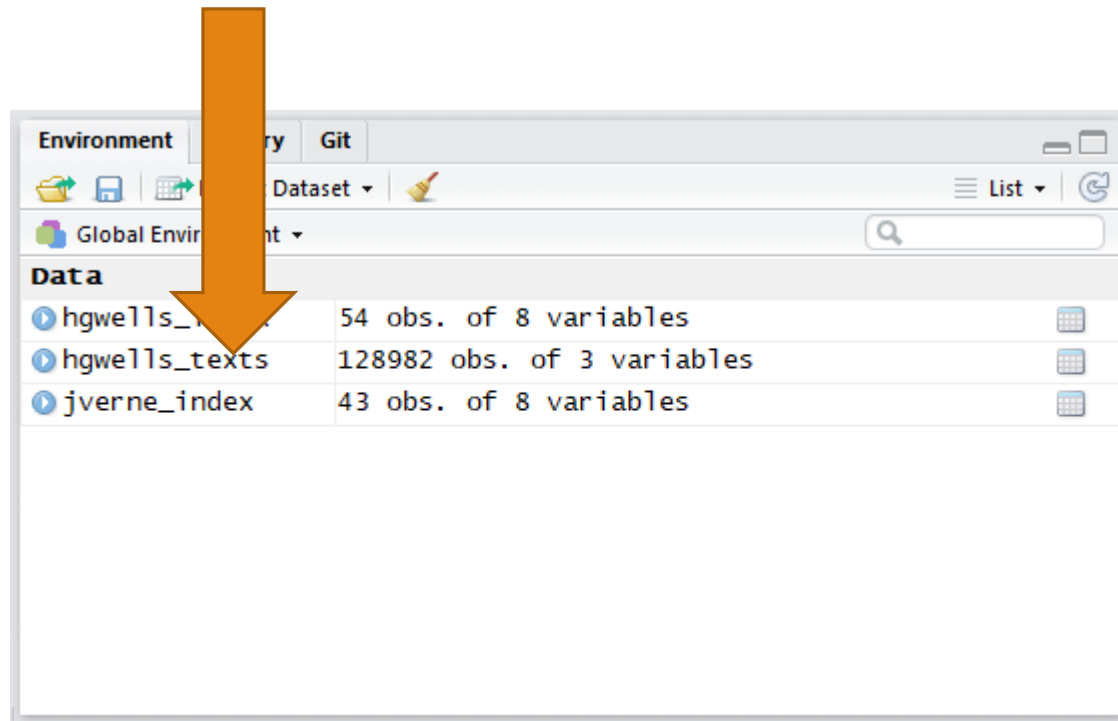
```
install_script.R * workshop1_gutenberg.R * gutenberg_metadata * hgwells_index * hgwells_ >>
66 filter(str_detect(author, "wells, H. G. ")) %>%
67 filter(str_detect(language, "en")) -> hgwells_index
68
69 gutenberg_metadata %>%
70   filter(has_text==TRUE) %>%
71   filter(str_detect(author, "Verne, Jules")) %>%
72   filter(str_detect(language, "en")) -> jverne_index
73
74 hgwells_texts <- gutenberg_download(hgwells_index$gutenberg_id[1:15],
75   meta_fields = "title")
76
77 jverne_texts <- gutenberg_download(jverne_index$gutenberg_id[1:15], meta_fiel
78   ds = "title")
79
80 #count (number of lines per book)
81 hgwells_texts %>%
82   count_lines()
83
84 jverne_texts %>%
85   count_lines()
86
87 #write to file
88 write_lines(hgwells_texts, "hgwells_texts.txt")
89 write_lines(jverne_texts, "jverne_texts.txt")
90
91 #read from file
92 read_lines("hgwells_texts.txt")
93 read_lines("jverne_texts.txt")
```

69:19 (Top Level) R Script

Data

To look at the data

To view a dataframe



How data looks like

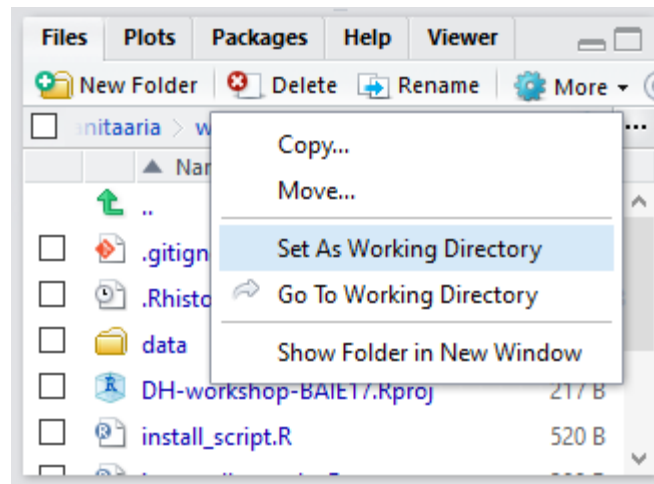
Just a table really 😊

	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	barbican	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

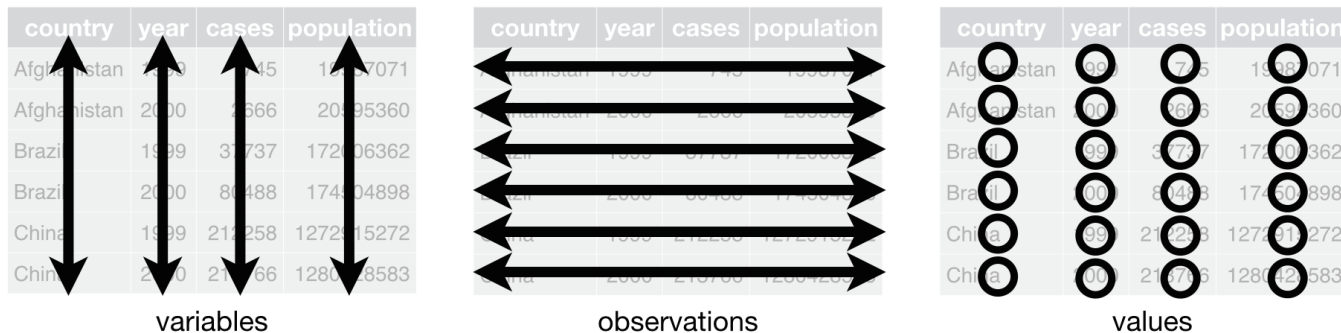
If you do not use Rproject:

- 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.



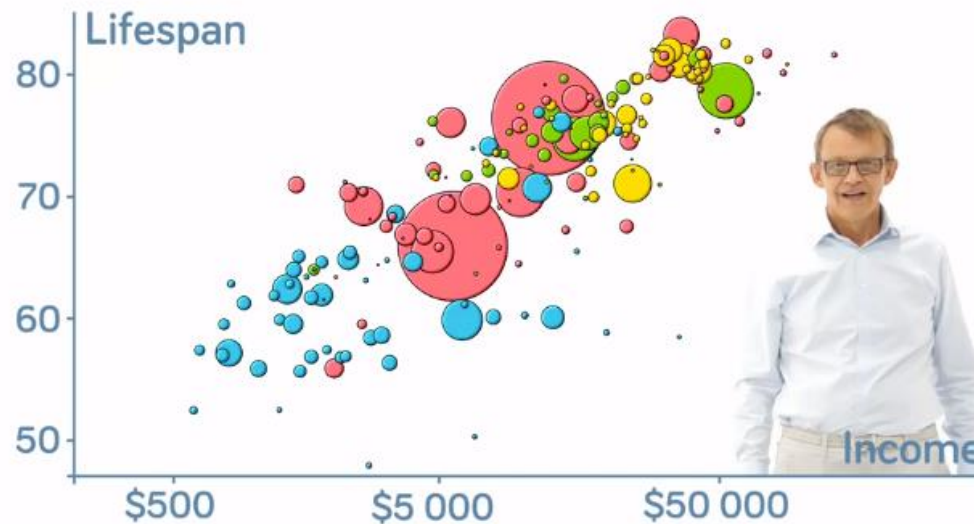
<http://vita.had.co.nz/papers/tidy-data.html>



dataset

How Does Income Relate to Life Expectancy?

Short answer - Rich people live longer



<https://www.gapminder.org/answers/how-does-income-relate-to-life-expectancy/>

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

<https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.html>

<http://style.tidyverse.org/>

<http://www.significantdigits.org/2017/10/switching-from-base-r-to-tidyverse/>

https://rpubs.com/bradleyboehmke/data_wrangling

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