

R ja tidyverse

LIHTNE TEKSTIANALÜÜS R-IS

PEETER TINITS 04.04.2019

Programmeerimiskeel R



R

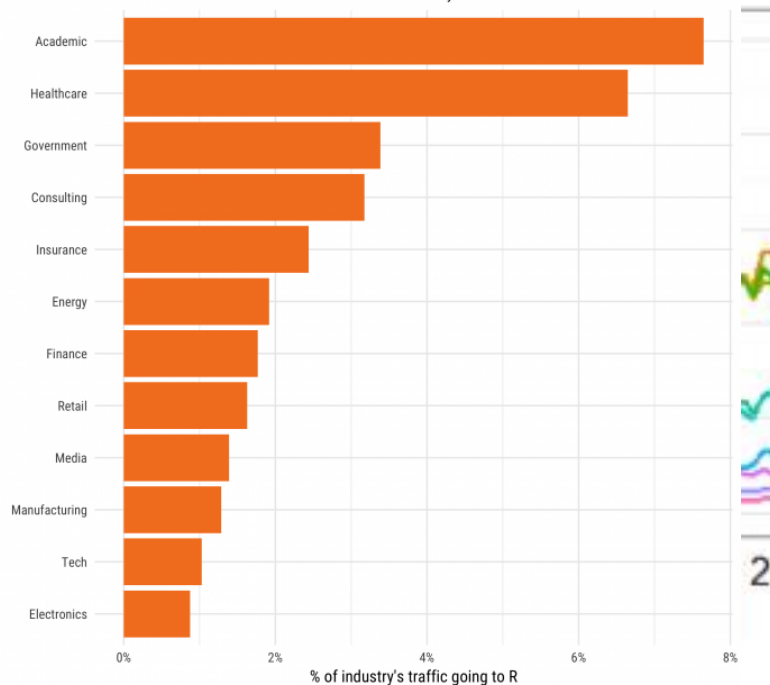
- Programmeerimiskeel ja vaba tarkvara keskkond
- Statistilisteks arvutuseks ja visualiseerimiseks.
- (R Foundation for Statistical Computing)
- R keelt kasutavad palju statistikud ja andmekaevandajad statistilise tarkvara ja andmeanalüüsi arendamiseks.

-Wikipedia

R-i kasutajad

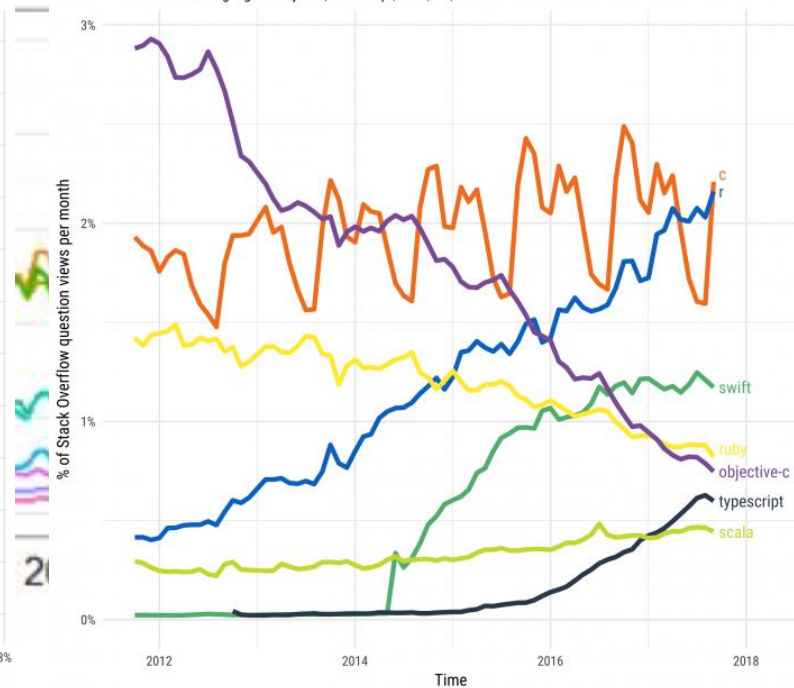
Visits to R by industry

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/>

Miks just R?

Avatud teadus ja selge analüüs

- Skriptifaili saab jagada
- Skriptifaili saab lugeda ja kontrollida

Kiirem, efektiivsem ja kindlam töö

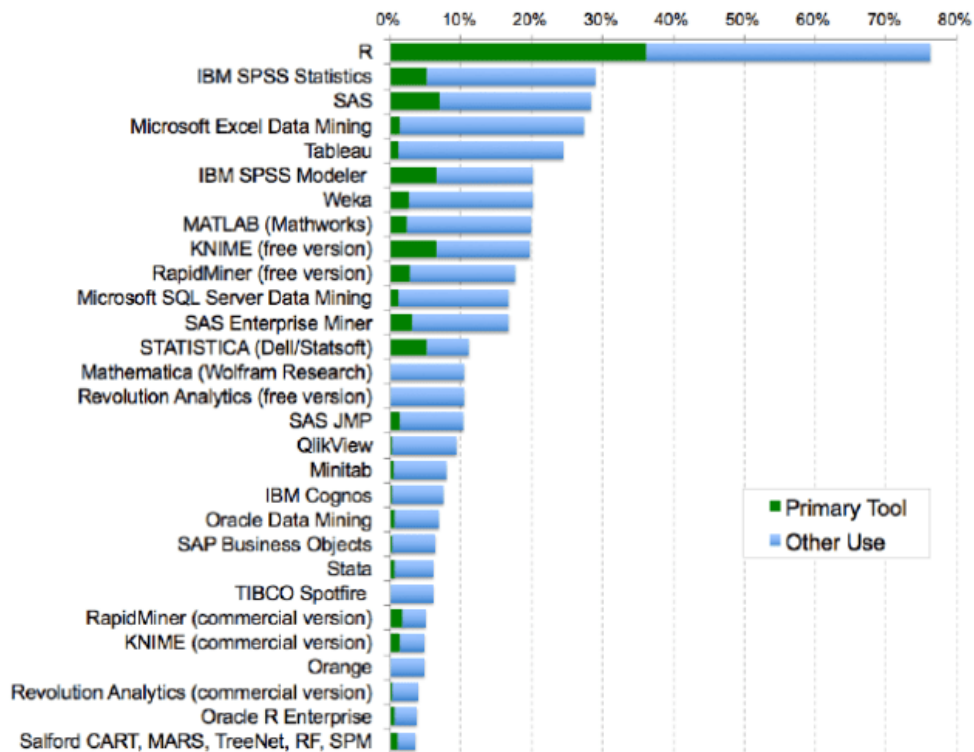
- Skriptifail dokumenteerib analüüsi
- Skriptifail jääb ja on taaskasutatav
- Kui R saab selgemaks saab ta ka mugavaks

Kogukond, mis toetab

- Pea lõputu arv pakette probleemide lahendamiseks
- Aktiivsed kogukonnad, kus on probleemid üldiselt juba läbi arutatud

Millist tarkvara kasutad

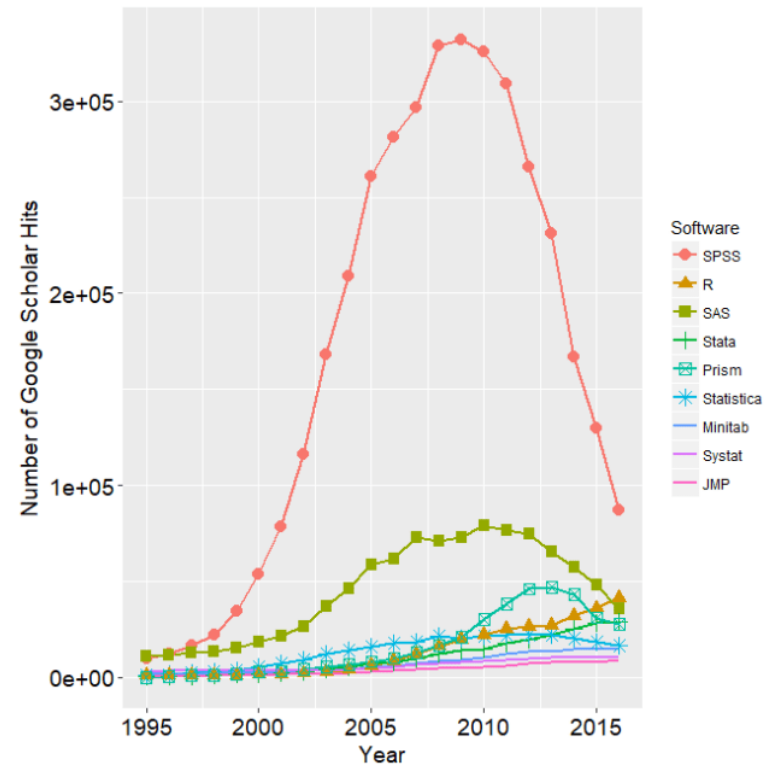
Andmeteadlaste küsitlus (n=1220), 2015



[Muenchen 2019. The Popularity of Data Science Software](#)

Andmeanalüüsi tarkvarad

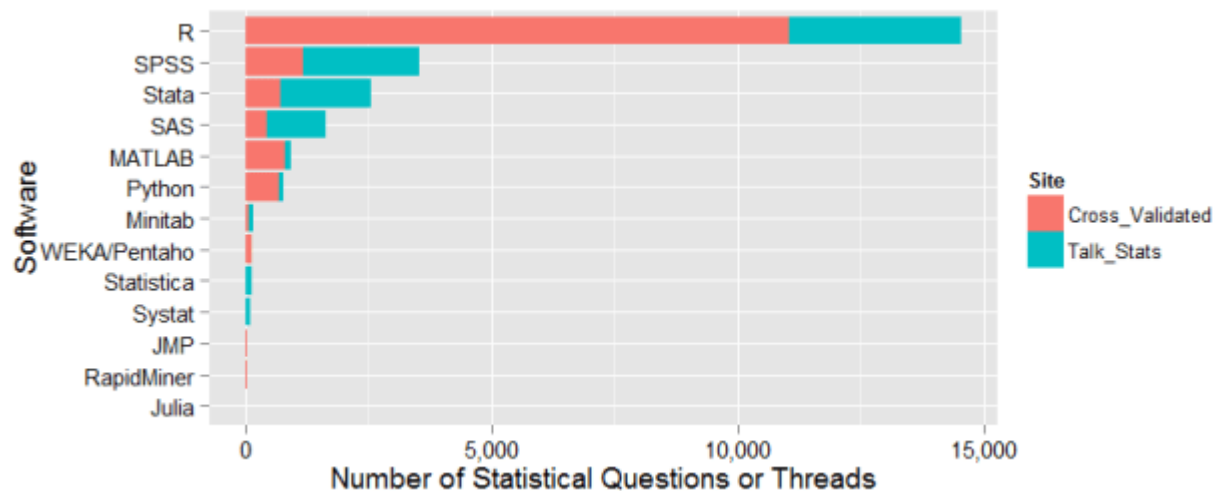
Viited akadeemilises kirjanduses



[Muenchen 2019. The Popularity of Data Science Software](#)

Kogukond

Küsimused statistikakogukondades [Talk Stats](#) ja [Cross Validated](#)

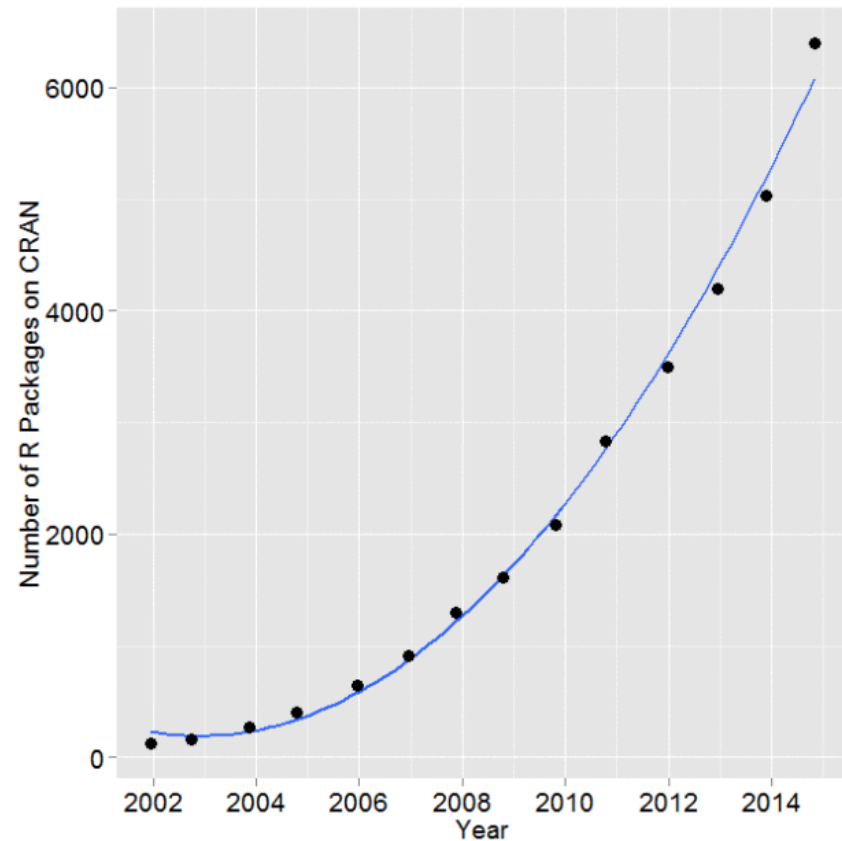


[Muenchen 2019. The Popularity of Data Science Software](#)

Pakettide arv

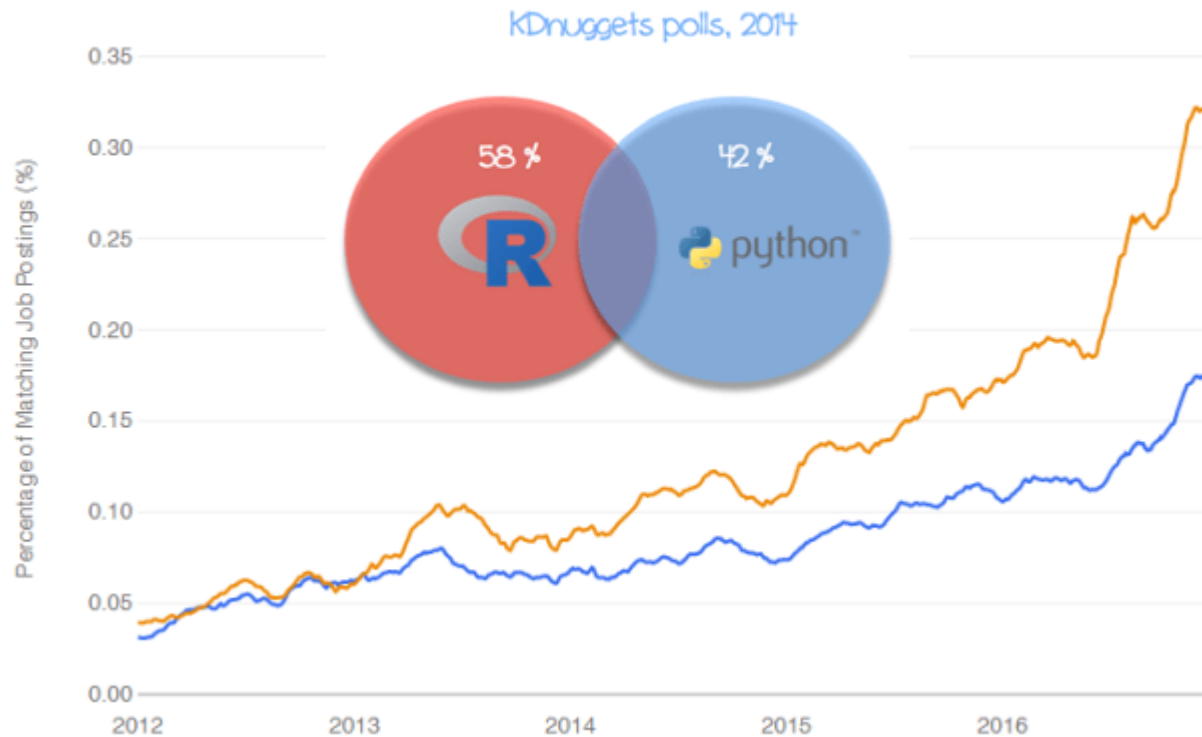
Pea eksponentsiaalne kasv.

Aastal 2015 rohkem uusi pakette kui SAS-is on kokku funktsioone



[Muenchen 2019. The Popularity of Data Science Software](#)

R vs Python



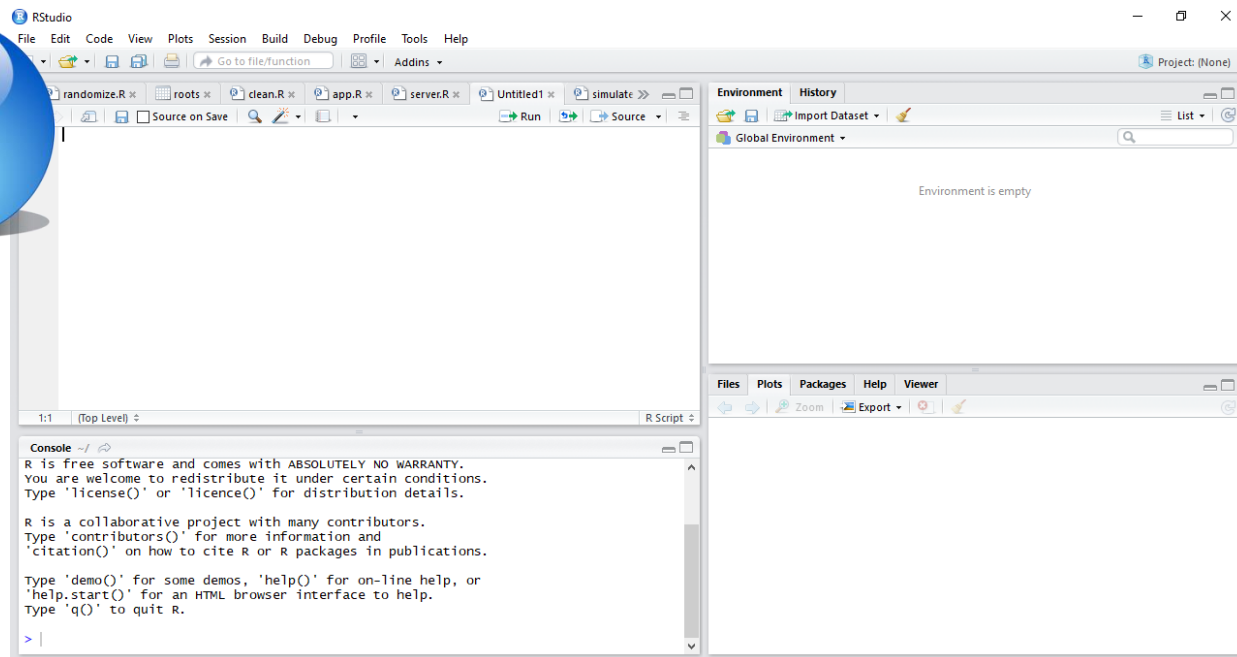
Töökohakirjeldused: R (sinine) ja Python (oranž).

[Muenchen 2019. The Popularity of Data Science Software](#)

RStudio

Rstudio – tasuta ja vaba lähte koodiga integreeritud arenduskeskkond

Sinu aken R-i!



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

Andmetega töötamine

Andmetega töötades pead Sa:

- Mõtleva välja, mida tahad teha.
- Kirjeldama neid ülesandeid programmeeriskeeles.
- Käivitama kirjutatud programmi.

Tidyverse teeb need sammud kiireks ja lihtsaks:

- Pakub lihtsaid ja arusaadavaid käske.
- Andmete töötlust tehakse loetavalt.
- Enamasti ka kiirem kui tava-R.

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

Kuidas kasutada R-i



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



Alustame R-iga

Lae alla andmed

Pealeht:

<https://github.com/peeter-t2/TM-TartuSpring2019>

Juhendid alla kerides:

1. Lae alla kõik failid.
2. Selleks vajuta download .zip lingile ülal paremal rohelise nupu all või mine siia: <https://github.com/peeter-t2/TM-TartuSpring2019/archive/master.zip>
3. Unpack the files where you want them.









Või kasuta Git-i, juhend sealsamas.

Alustuseks

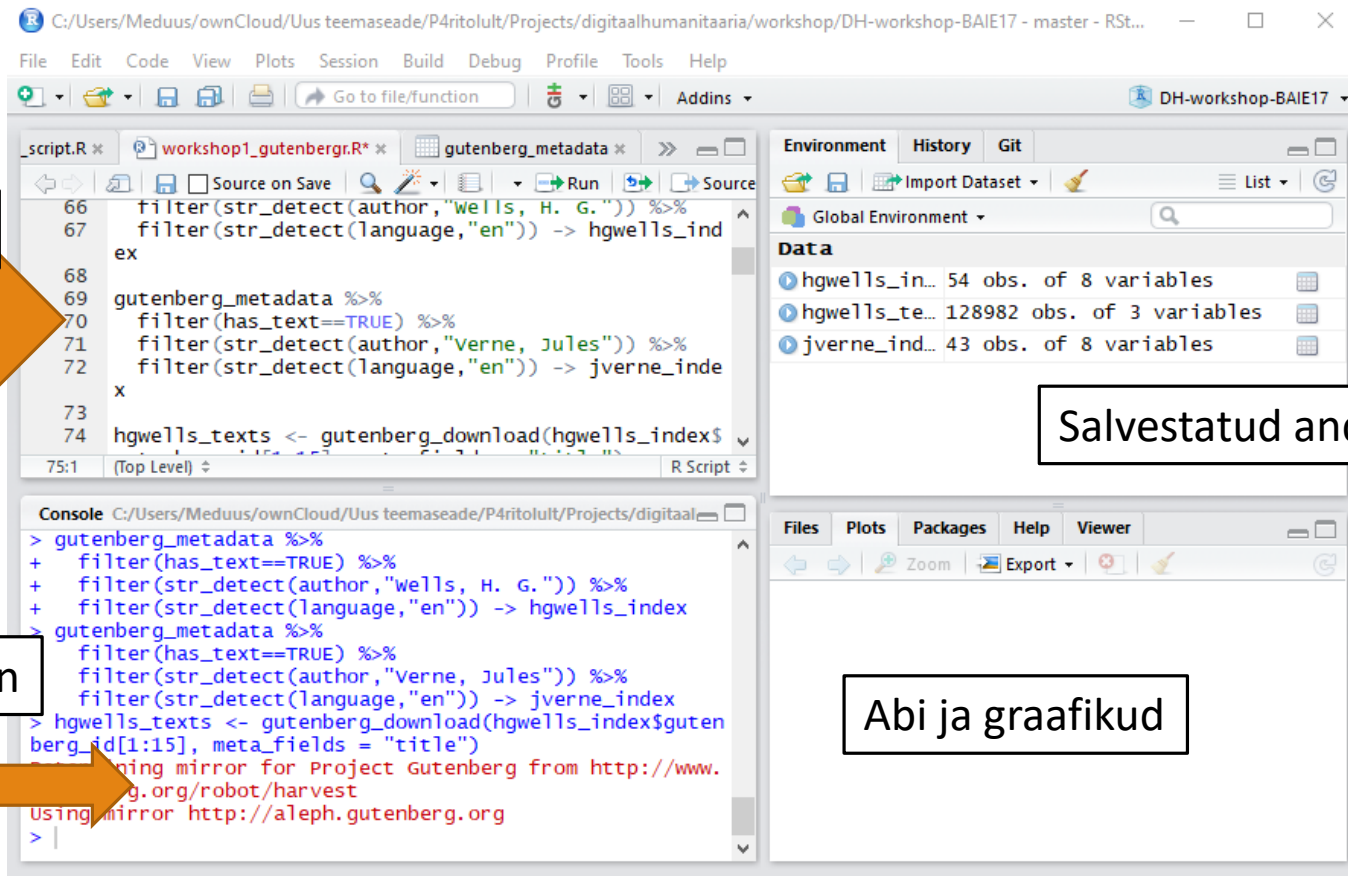
Alustuseks jookсутa Rproj faili. Selline algus lihtsustab natuke Rstudio kasutamist, seades automaatselt töökataloogi ja pidades meeles mõningaid seadeid.



Name

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

RStudio vaade



Skripti failid

Roheline = kommentaarid, (ja tekstiandmed – nt „Wells, H. G.“)

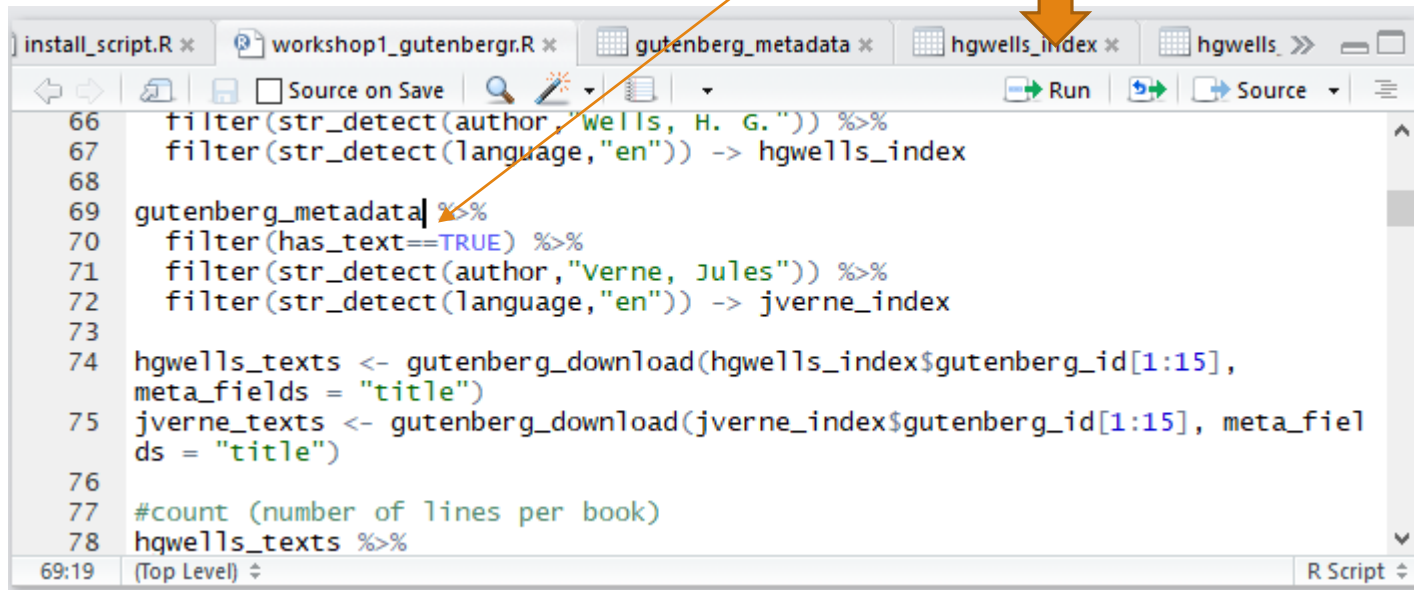
must, sinine, jne = käsud

```
☐ 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 ↕
```

R-i jooksutamine

To run code

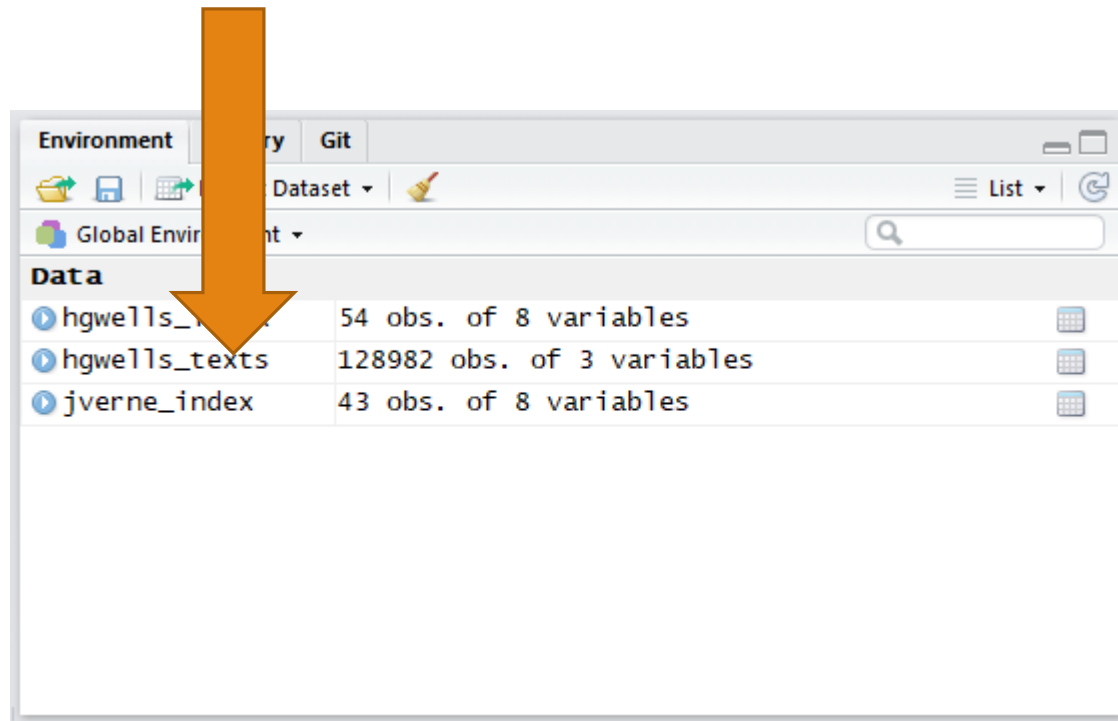
Pick a line and click run



Andmed

Andmete vaatamiseks

To view a dataframe



Kuidas andmed välja näevad

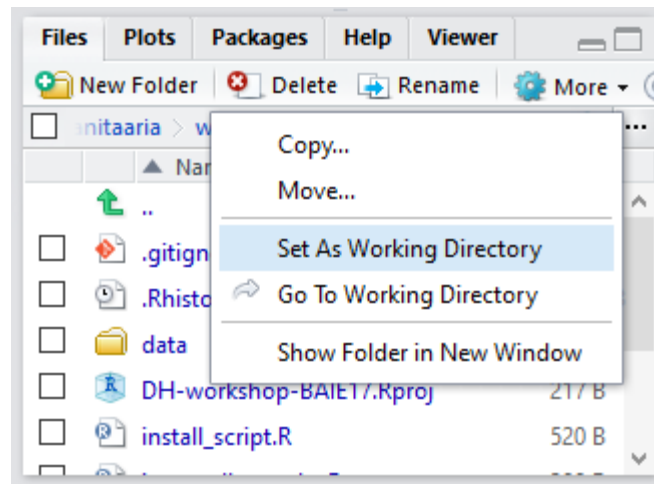
Ka tekstiandmed on lihtsalt tabel

	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

Failide kasutamiseks

Kui sa ei kasuta Rprojeki:

- Pead sättingima alustuseks õige töökataloogi



Tidy data / Puhtad andmed

Tidy/Puhtad andmed on lihtsasti töödeldavad, modelleeritavad ja visualiseeritavad. Neil on kindel struktuur: iga tulp on muutuja, iga vaatluskord on rida ja iga vaatluskogum on tabel.

country	year	cases	population
Afghanistan	1999	31745	15467071
Afghanistan	2000	2666	20495360
Brazil	1999	31737	172006362
Brazil	2000	80488	174604898
China	1999	212258	1272015272
China	2000	216766	1280423583

variables

country	year	cases	population
Afghanistan	1999	31745	15467071
Afghanistan	2000	2666	20495360
Brazil	1999	31737	172006362
Brazil	2000	80488	174604898
China	1999	212258	1272015272
China	2000	216766	1280423583

observations

country	year	cases	population
Afghanistan	99	75	15467071
Afghanistan	00	66	20495360
Brazil	99	737	172006362
Brazil	00	488	174604898
China	99	2258	1272015272
China	00	6766	1280423583

values

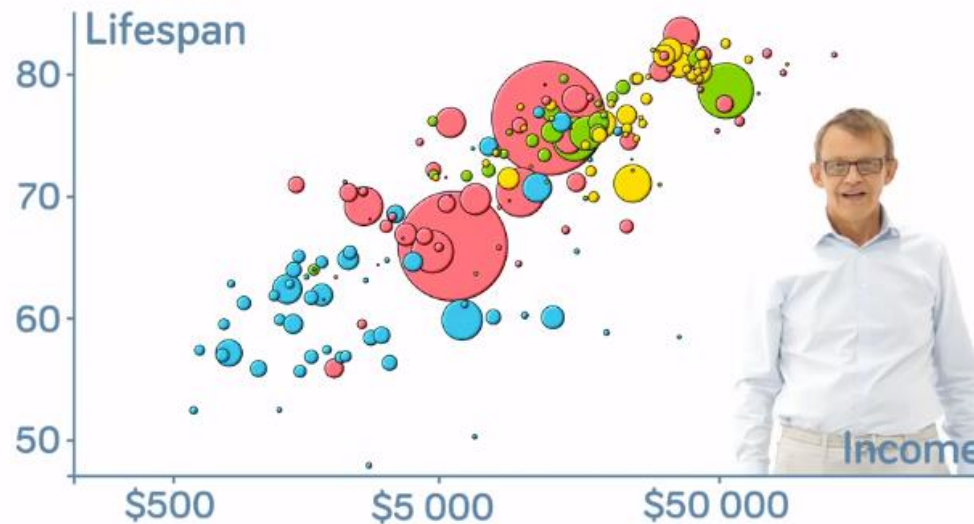
<http://vita.had.co.nz/papers/tidy-data.html>



andmestik

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/>

Lihtsamad tidyverse käsud

%>% - saadab objekti töötlusesse järgmisel real

andmed %>%

protsess()

select() – vali kindel muutuja

filter() – filtreeri osad andmed välja

group_by() – grupeeri andmed mingite kategooriate kaupa

summarise() – võtta andmed mõne funktsiooniga kokku

arrange() – järjestada andmed

join() – ühenda tabelid omavahel

mutate() – tee uus muutuja

Tidyverse allikaid

<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

Lisa: Mitte-tidyverse R 1

Pesastatud võimalus:

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

https://rpubs.com/bradleyboehmke/data_wrangling

Lisa: Mitte-tidyverse R 2

Mitme objekti võimalus:

```
a <- filter(data, variable == numeric_value)
```

```
b <- summarise(a, Total = sum(variable))
```

```
c <- arrange(b, desc(Total))
```

https://rpubs.com/bradleyboehmke/data_wrangling

Lisa: Tidyverse R

%>% võimalus:

data %>%

filter(variable == "value") %>%

summarise(Total = sum(variable)) %>%

arrange(desc(Total))

https://rpubs.com/bradleyboehmke/data_wrangling