The unexpected value of R in Official Statistics

Edwin de Jonge

e.dejonge@cbs.nl

Let’s start with a puzzling remark

In statistics production, **most time** and effort is spend **in other activities** than statistical analysis.

- Statistics education mostly deals with statistical analysis and methods
- R excels in applying statistical methods, does that limit its application?

Who am I?

- Edwin de Jonge
- > 20 yrs at Statistics Netherlands (CBS)
- Methodologist / Data Scientist
- Working a lot with R, mainly...

Expertise: methods in data viz, cleaning and network analysis. Most of them in R

What are “Official Statistics”

Official statistics are statistics published by government agencies or other public bodies such as international organizations as a public good.

Wikipedia, 2021

Official Statistics as a public good

- Statistics available for **citizens**, **society** and **science**
- Methods must be documented
- Production should be transparent / cost efficient

R and R packages are a **natural fit**:

- statistical **methods**/tools are **open** and documented.
- **R packages** can be **shared** between offices

What are “Official Statistics”

Official statistics are ***statistics*** *published by government agencies or other public bodies such as international organizations as a public good.*

Wikipedia, 2021

Puzzling remark

In statistics production, most time and effort is spend in other activities than statistical analysis.

What other activities?

What is statistics?

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data

Wikipedia 2021

What is statistics?

*Statistics is the discipline that concerns the **collection**, **organization**, **analysis**, **interpretation**, and **presentation** of data*

Wikipedia 2021

Official Statistics as a public good

- **Statistics available for citizens, society and science**

Public good:

So official statistics = statistics activities + **making available**

Many R packages are used to disseminate/publish statistics

How R supports Official Statistics

Let’s look at these sources:

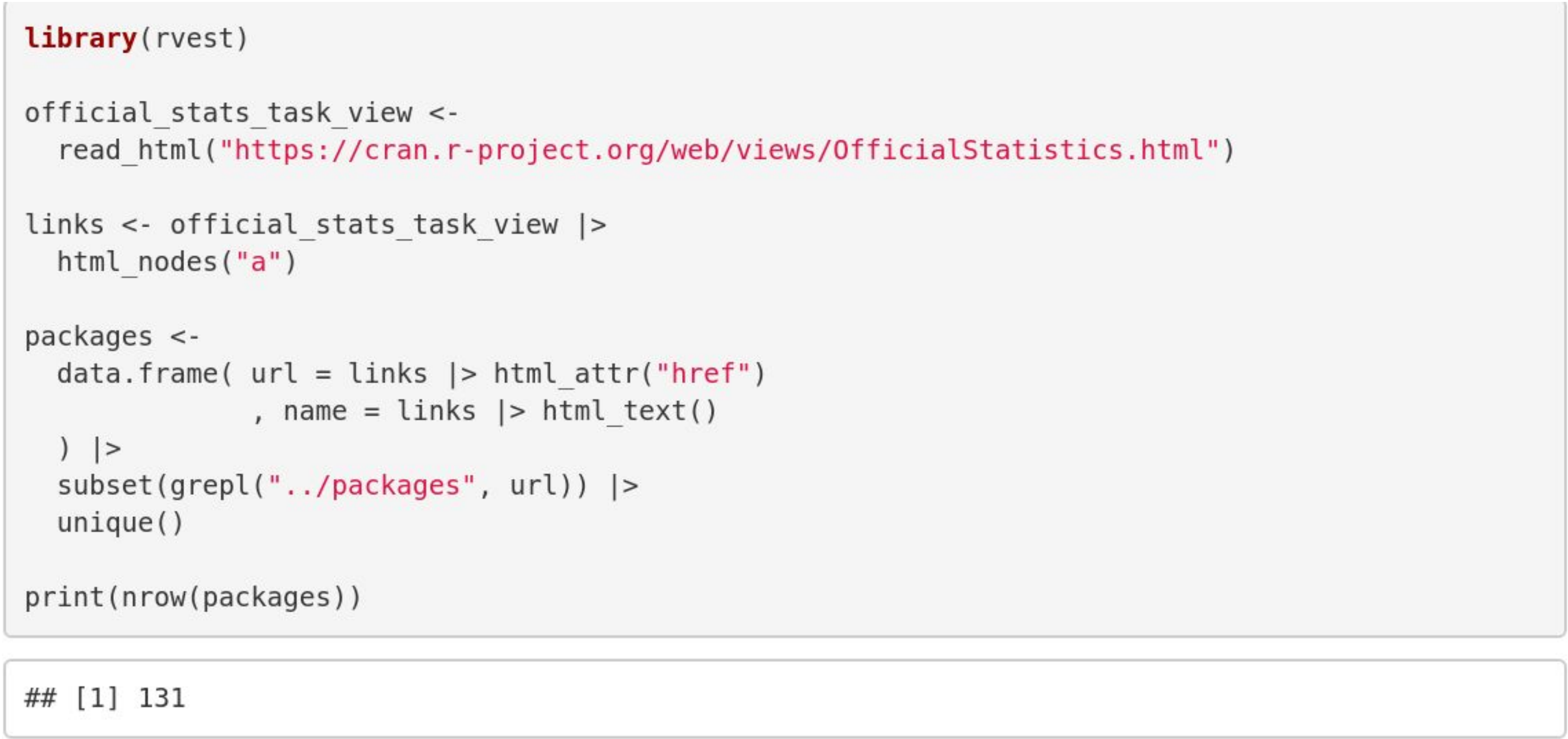
- CRAN Task View: Official Statistics & Survey Statistics (M. Temple)
- Awesome official statistics software (O. Bosch et al.)

CRAN Task View Official Stats

Topics:

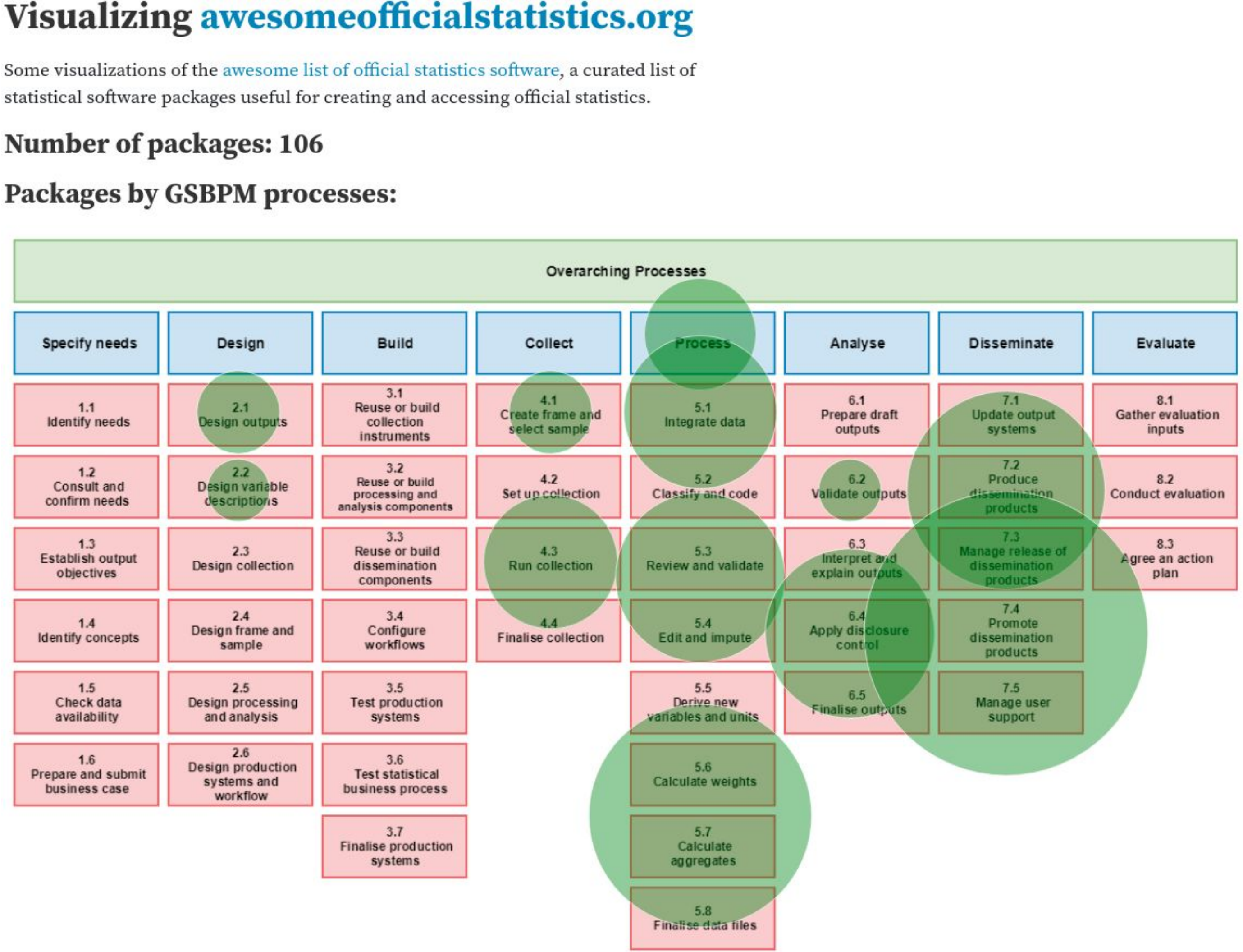
- Complex Survey Design: Sampling and Sample Size Calculation
- Complex Survey Design: Point and Variance Estimation and Model Fitting
- Complex Survey Design: Calibration
- Editing and Visual Inspection of Microdata
- Imputation
- Statistical Disclosure Control
- Seasonal Adjustment and Forecasting
- Statistical Matching and Record Linkage
- Small Area Estimation
- Indices, Indicators, Tables and Visualisation of Indicators
- Microsimulation and synthetic data
- Additional Packages and Functionalities

CRAN Task View Official Stats: 131 packages



```r
library(rvest)

official_stats_task_view <-
  read_html("https://cran.r-project.org/web/views/OfficialStatistics.html")

links <- official_stats_task_view |>
  html_nodes("a")

packages <-
  data.frame( url = links |> html_attr("href")
            , name = links |> html_text()
  ) |>
  subset(grepl("../packages", url)) |>
  unique()

print(nrow(packages))
```

```
## [1] 131
```



Visualizing awesomeofficialstatistics.org

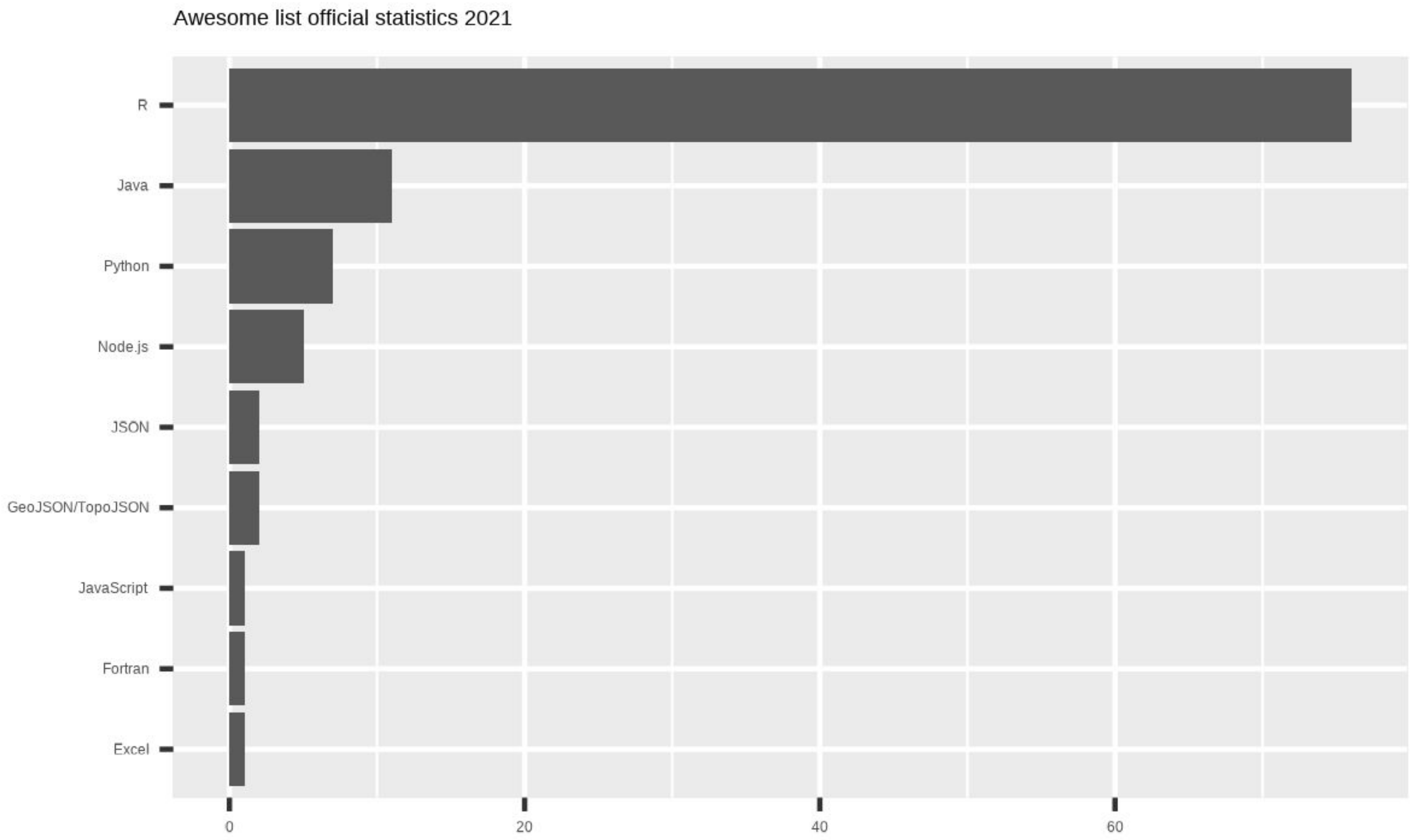
Some visualizations of the awesome list of official statistics software, a curated list of statistical software packages useful for creating and accessing official statistics.

Number of packages: 106

Packages by GSBPM processes:

Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			


Awesome list official statistics 2021
R
Java
Python
Node.js
JSON
GeoJSON/TopoJSON
JavaScript
Fortran
Excel
0
20
40
60

So R seems useful for Official Statistics

- 72% of awesome software is R!
- many unexpected uses outside the realm of statistical estimation.

Other arguments/practices: *Use of R for Official Statistics*, Kowarik, van der Loo, 2018

Unexpected value of R

R gives freedom to develop what you need

- brings you into a wide array of applications: i.e. not just statistical estimation, all other parts of the statistical process, including publication.
- R helps tremendously in filling the gaps in the statistical process, which is unexpected
- We all know and use that (e.g. data.table, tidyverse, rmarkdown, shiny, etc.)

Unexpected value of R

R brought me in unexpected places:

a.o.:

- Visual validation
- data cleaning
- statistical disclosure control
- dashboards

Visual validation

The main task of graphics are:

- *to reveal the unexpected*
- *to make the complex easier to perceive*

John Tukey

Tableplot as data quality tool

- Visual summary of large dataset
- can be used to assess quality aspects / distributions of large datasets.

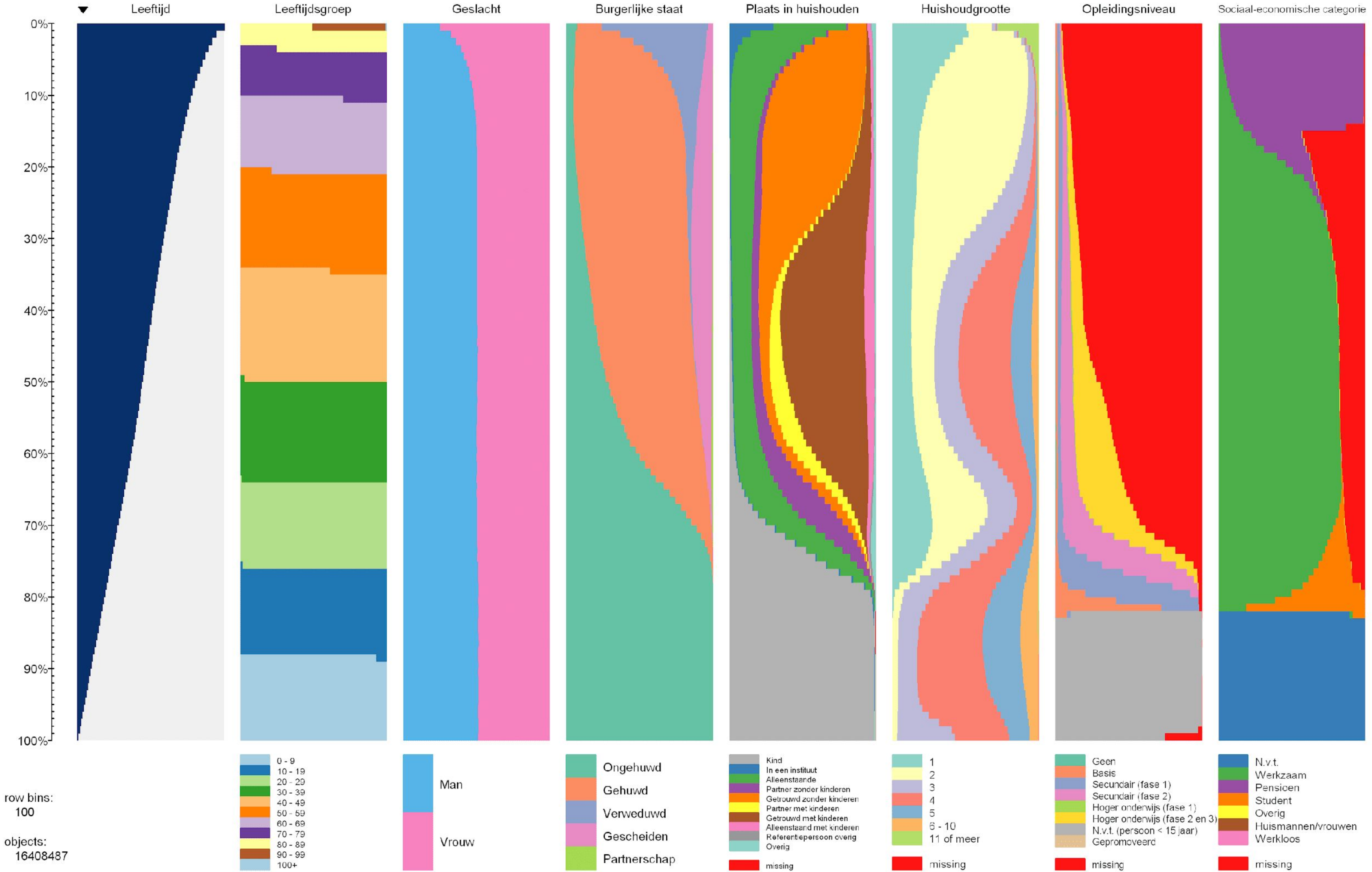
Visualizing and Inspecting Large Datasets with Tableplots, Journal of Datascience 2013, M. Tennekes, E. de Jonge, P. Daas

Tableplot recipe

- Take a dataset, preferable large
- Sort it on a numerical variable, e.g. “age”
- Chop the dataset in 100 slices
- Calculate for each slice the frequency (or mean) for each variable.
- Plot these and inspect
- Enjoy!

Case: Dutch Census

- NL has a virtual census, compiled with population registers
- combination of these registers needs some checking
- dataset is 17M persons



Leeftijd
Leeftijdsgroep
Geslacht
Burgerlijke staat
Plaats in huishouden
Huishoudgrootte
Opleidingsniveau
Sociaal-economische categorie
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
row bins:
100
objects:
16408487
0 - 9
10 - 19
20 - 29
30 - 39
40 - 49
50 - 59
60 - 69
70 - 79
80 - 89
90 - 99
100+
Man
Vrouw
Ongehuwd
Gehuwd
Verweduwd
Gescheiden
Partnerschap
Kind
In een instituut
Alleenstaande
Partner zonder kinderen
Getrouwd zonder kinderen
Partner met kinderen
Getrouwd met kinderen
Alleenstaand met kinderen
Referentiepersoon overig
Overig
missing
1
2
3
4
5
6 - 10
11 of meer
missing
Geen
Basis
Secundair (fase 1)
Secundair (fase 2)
Hoger onderwijs (fase 1)
Hoger onderwijs (fase 2 en 3)
N.v.t. (persoon < 15 jaar)
Gepromoveerd
missing
N.v.t.
Werkzaam
Pensioen
Student
Overig
Huismannen/vrouwen
Werkloos
missing

Data-cleaning packages

Or R as a Domain Specific Language for data validation

Datacleaning

NSI’s spend a lot of time “fixing” their data:

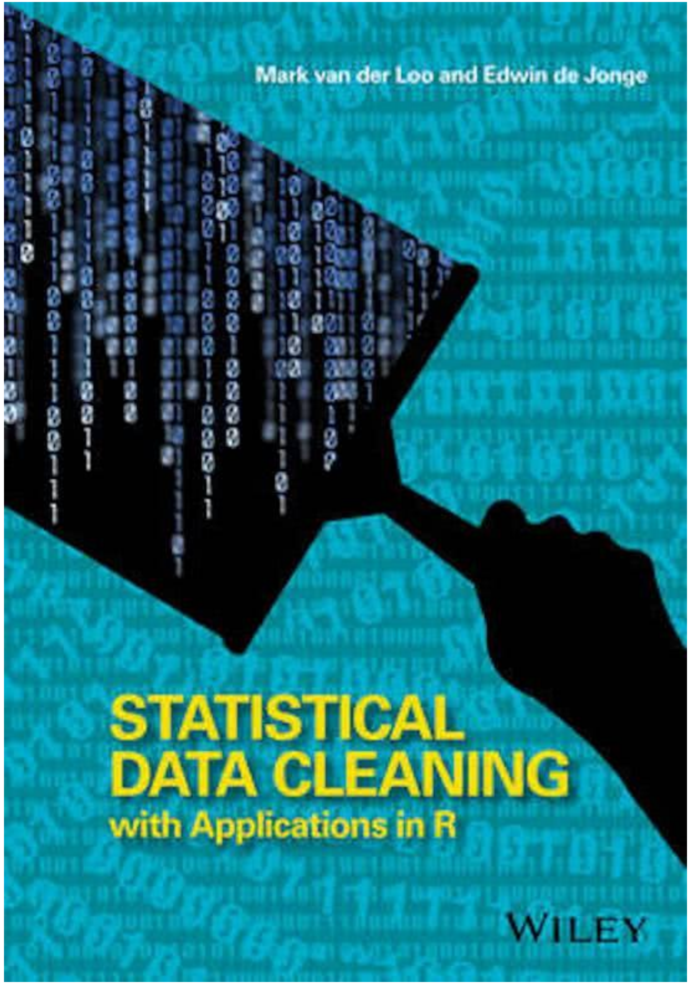
- checking the quality
- removing errors/outliers
- “correcting”/adjusting data.
- imputing missing data

R helps making this process, reproducible and transparent

Data-cleaning packages

Many R packages can be found:

- VIM, mice (imputation), simputation
- datamaid
- **validate**, **validatetools**
- **errorlocate**
- **dcmodify**



Coauthor data-cleaning package

- validate, validatedb, validatetools
- errorlocate
- validatesuggest
- dcmodify, dcmodifydb

All have the same underlying principle:

RULES

Use R to specify rules

- Use R syntax to specify a rule
- Store, annotate and document the rules
- When new data arrives, evaluate/execute the rules in the context of data.

Or “confront the data with the rules”

R can express itself

R is a powerful language:

- all language constructs are “expressions” (not statements)
- homo-iconic: it can interpret, modify, create and evaluate its own expressions.
- Few language can do that: LISP, julia
- it’s evaluation is very powerful: access to all scopes/environments

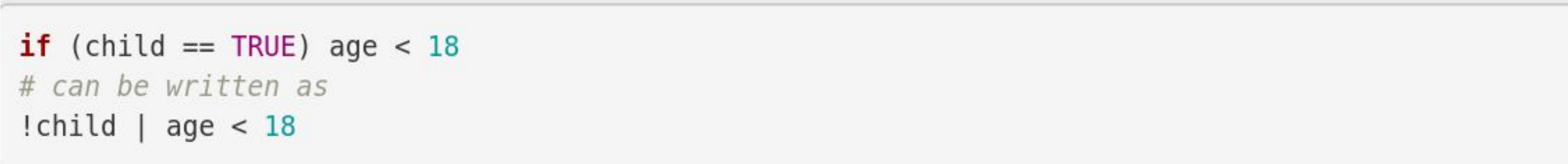
Used by key R packages

- formulas: y ~ x are expressions, evaluated in de context of a data.frame
- plot / ggplot use this to specify data mappings and titles
- data.table: dat[, child := age < 18]
- dplyr: mutate(dat, child = age < 18]

functions: quote, substitute, bquote, eval, rlang

All use expressions internally!

Use R to rewrite/optimize rule statement

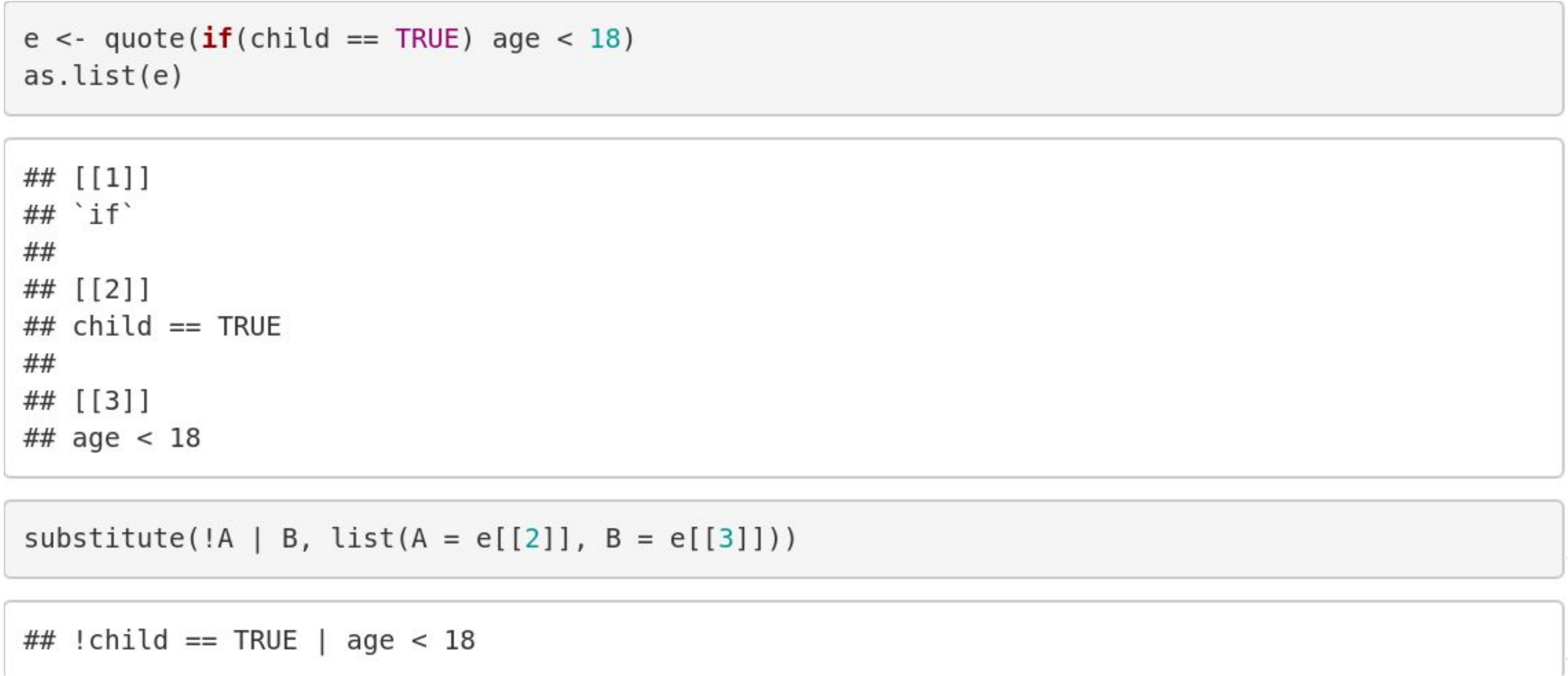


```r
if (child == TRUE) age < 18
# can be written as
!child | age < 18
```

First statement is familiar

Second can be used in a data.frame

R: quote, substitute



```r
e <- quote(if(child == TRUE) age < 18)
as.list(e)
```

```
## [[1]]
## `if`
##
## [[2]]
## child == TRUE
##
## [[3]]
## age < 18
```

```r
substitute(!A | B, list(A = e[[2]], B = e[[3]]))
```

```
## !child == TRUE | age < 18
```

Rewriting R rules

Used in:

- **validate** to optimize for data.frame
- **validatedb** to support SQL generation
- **errorlocate** to translate it into a Mixed Integer Problem (MIP)
- **dcmodify** to create modification statements

Privacy-protecting maps

Privacy
shift

R-package sdcSpatial

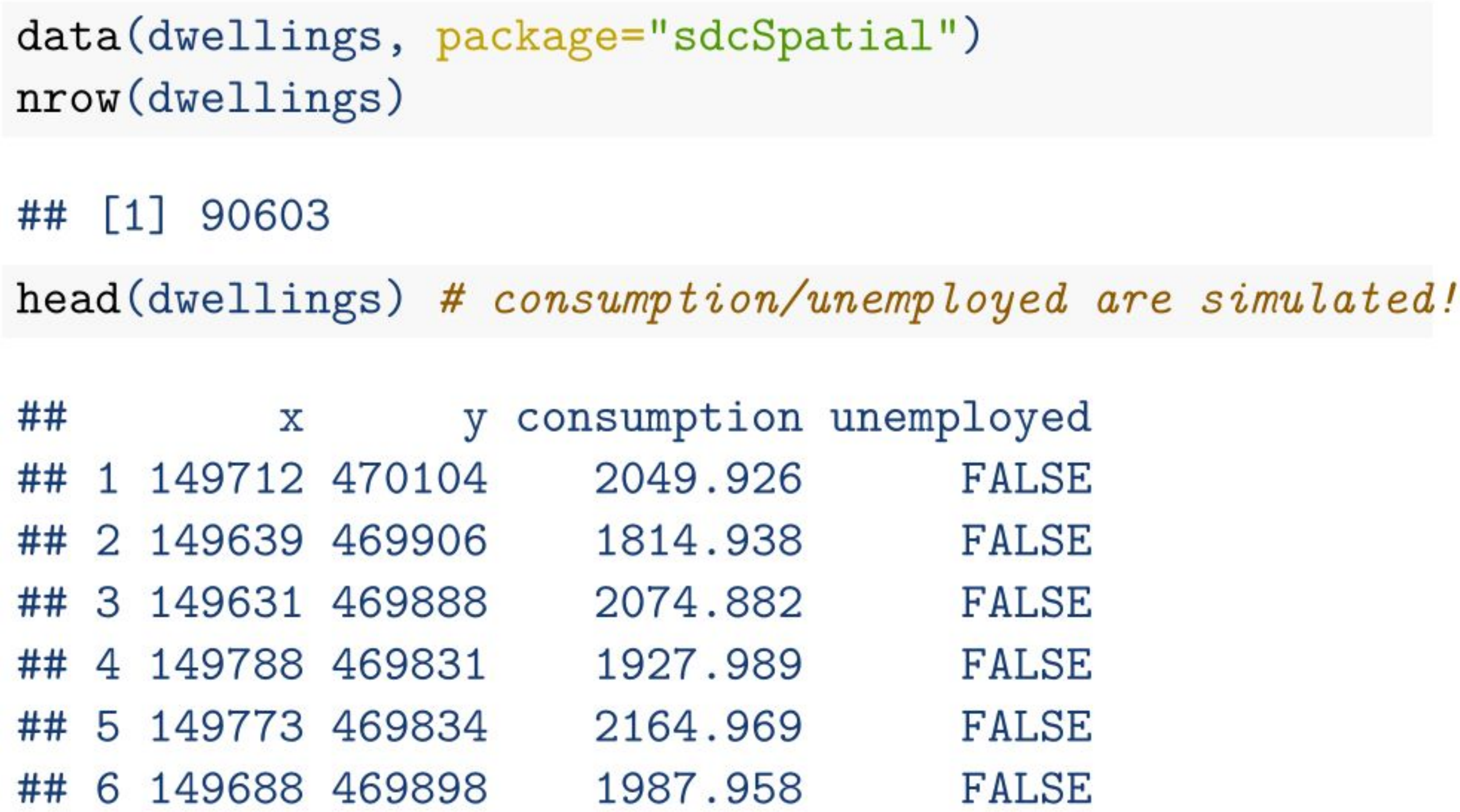
Idea:

Create a detailed density map without disclosing details of an individual.

Wolf, Peter-Paul de, and Edwin de Jonge. 2018. "Spatial Smoothing and Statistical Disclosure Control." In Privacy in Statistical Databases - PSD 2018

sdcSpatial

- works upon locations (e.g. addresses)
- aggregates these into a grid
- allows to (visually) assess the risk for disclosure
- contains protection methods, e.g. `protect_smooth`

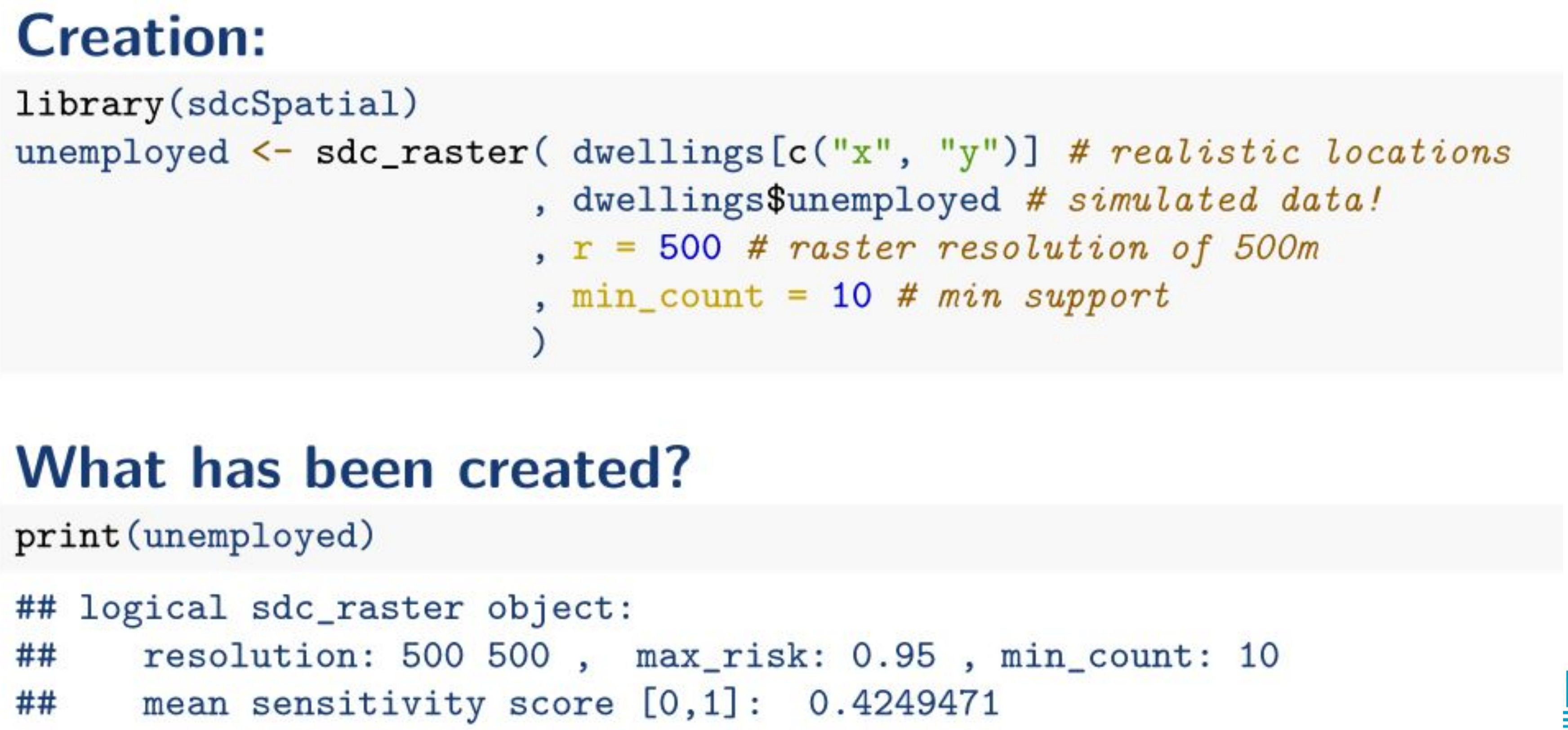


```
data(dwellings, package="sdcSpatial")
nrow(dwellings)
```

```
## [1] 90603
```

```
head(dwellings) # consumption/unemployed are simulated!
```

```
##        x      y consumption unemployed
## 1 149712 470104    2049.926      FALSE
## 2 149639 469906    1814.938      FALSE
## 3 149631 469888    2074.882      FALSE
## 4 149788 469831    1927.989      FALSE
## 5 149773 469834    2164.969      FALSE
## 6 149688 469898    1987.958      FALSE
```



Creation:

```
library(sdcSpatial)
unemployed <- sdc_raster( dwellings[c("x", "y")] # realistic locations
                        , dwellings$unemployed # simulated data!
                        , r = 500 # raster resolution of 500m
                        , min_count = 10 # min support
                        )
```

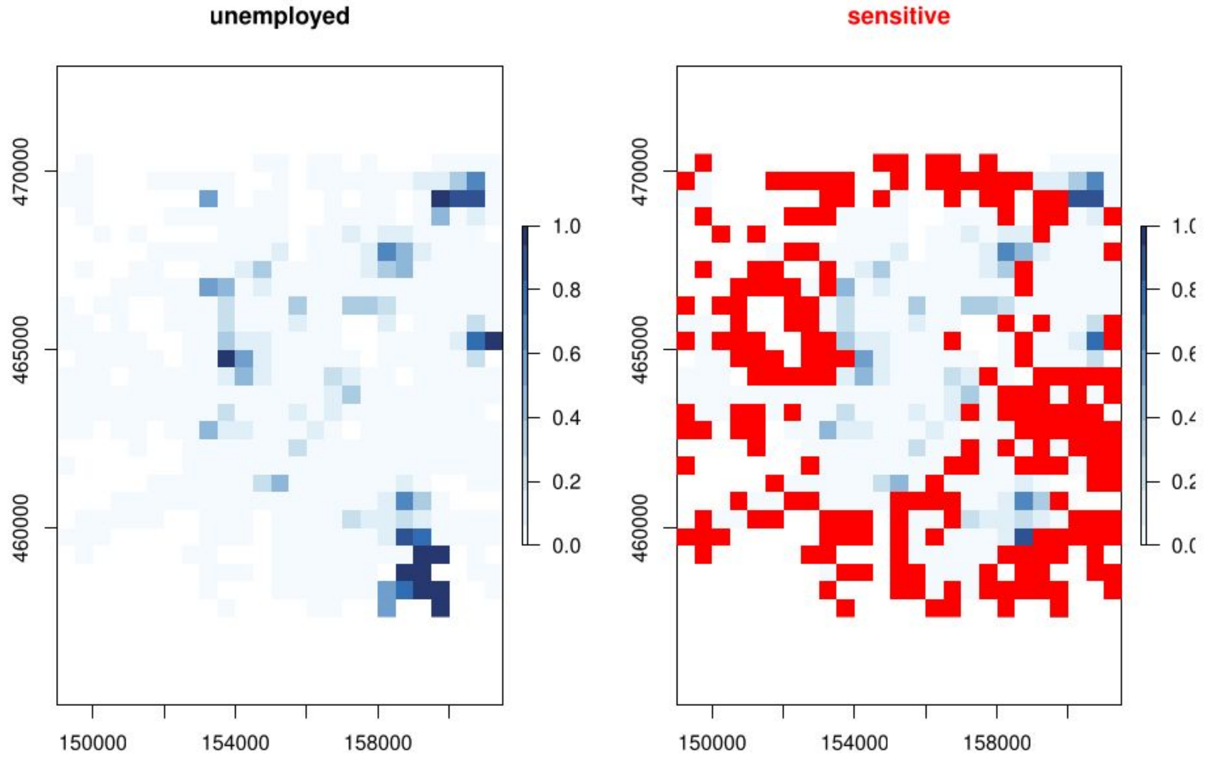
What has been created?

```
print(unemployed)
```

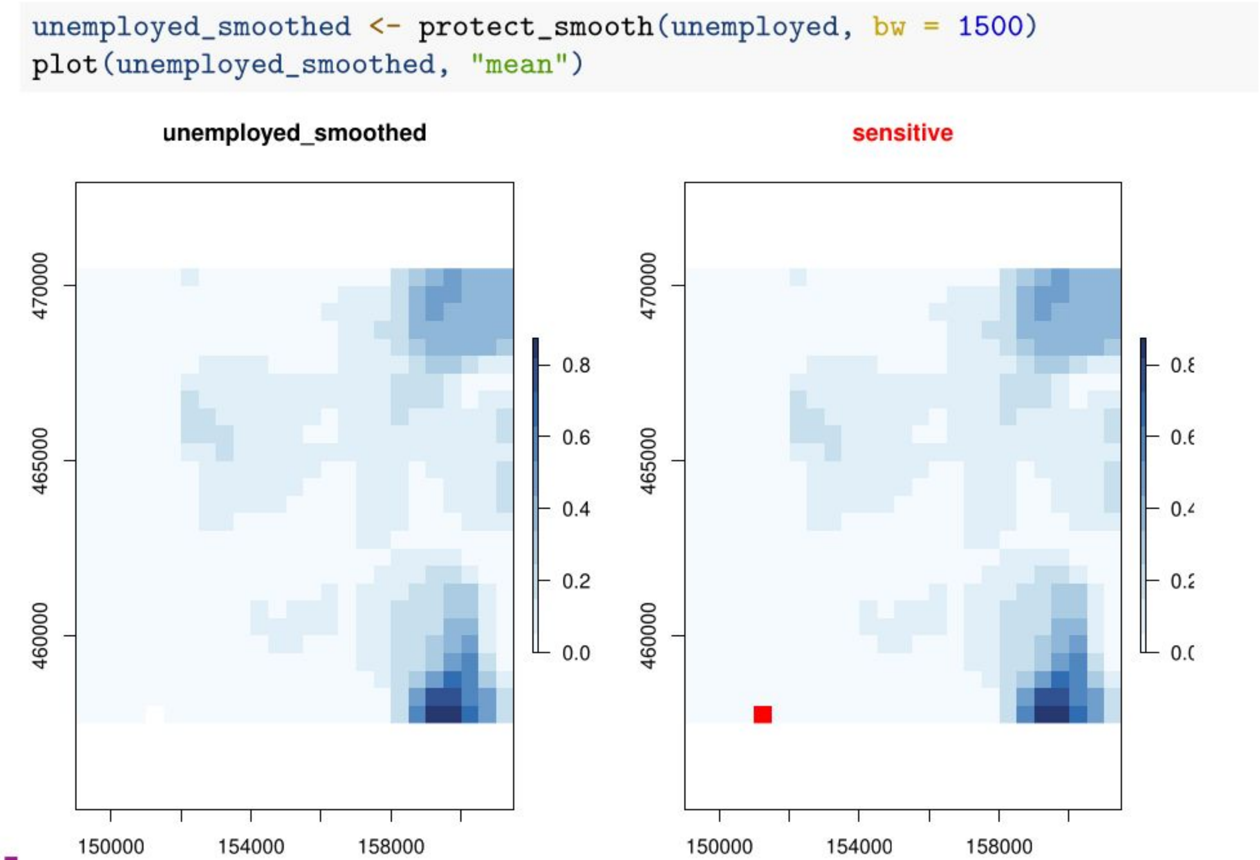
```
## logical sdc_raster object:
##    resolution: 500 500 ,  max_risk: 0.95 , min_count: 10
##    mean sensitivity score [0,1]:  0.4249471
```

Visual assess risk

```r
plot(unemployed, "mean")
```



And protect

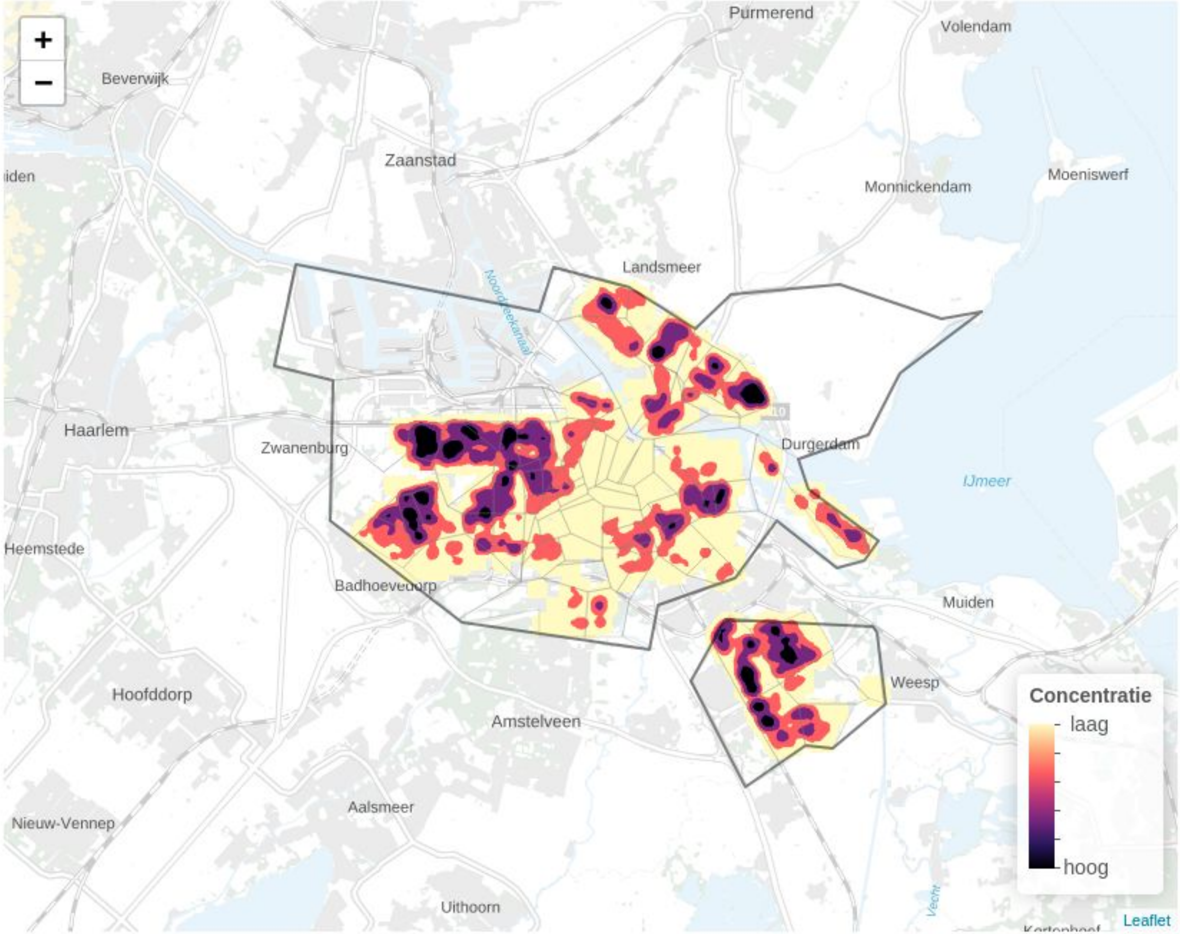


```
unemployed_smoothed <- protect_smooth(unemployed, bw = 1500)
plot(unemployed_smoothed, "mean")
```

Case: plotting educational disadvantages

- sdcSpatial used (offline) for plotting educational disadvantage maps https://dashboards.cbs.nl/v3/onderwijsachterstanden/
- as a leaflet layerd
- shiny dashboard
- Used by education department and municipalities to visualize high density locations.

Kaart

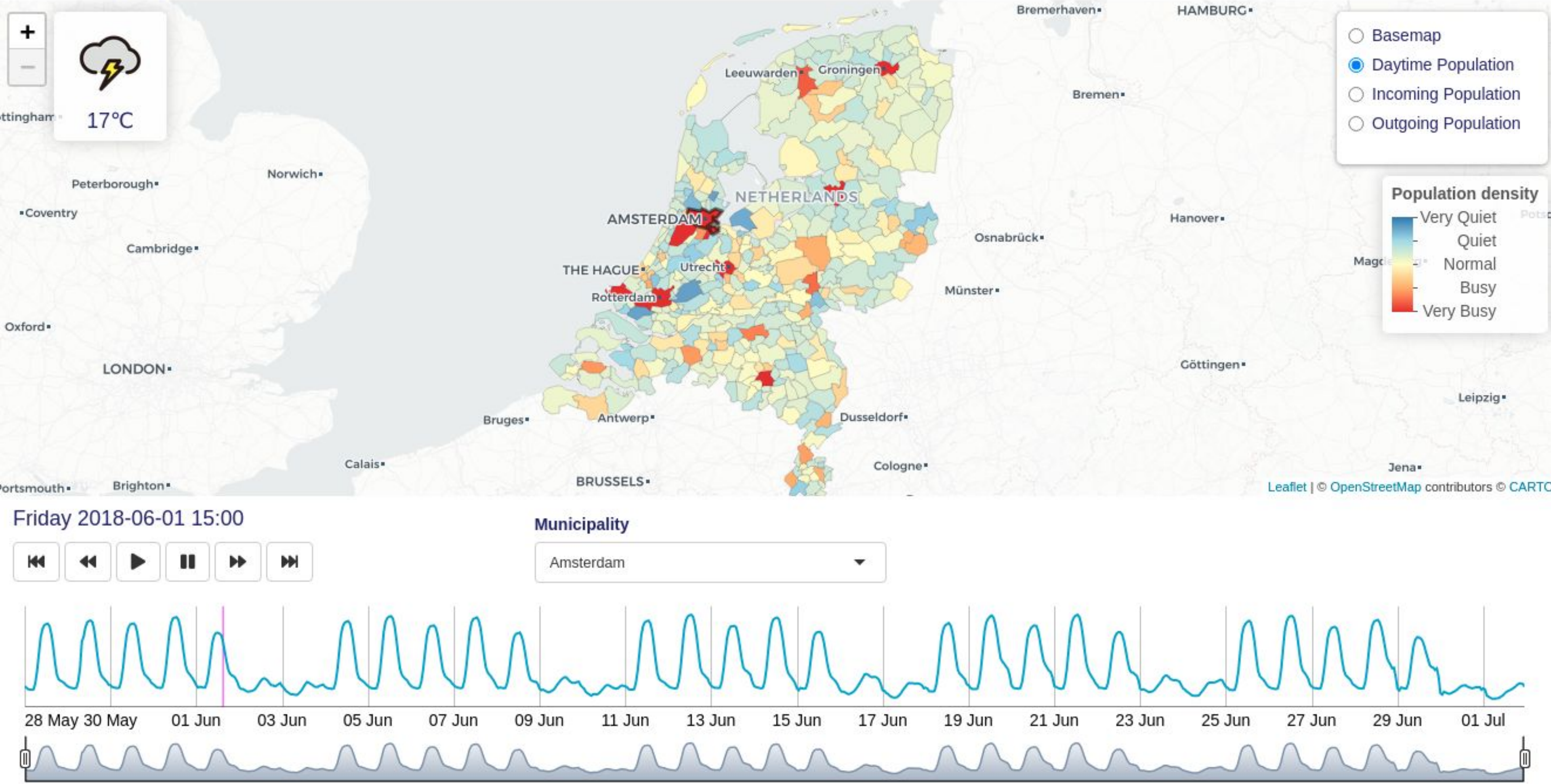

+
−
Purmerend
Volendam
Beverwijk
Zaanstad
iden
Monnickendam
Moeniswerf
Landsmeer
Noordzeekanaal
Haarlem
Zwanenburg
Durgerdam
IJmeer
Heemstede
Badhoevedorp
Muiden
Weesp
Hoofddorp
Amstelveen
Aalsmeer
Nieuw-Vennep
Uithoorn
Vecht
Kortenhoef
Leaflet
Concentratie
laag
hoog

Dashboards

Shiny dashboards

- At Statistics NL, a popular product for government departments
- Currently serving over 50 dashboards for external use
- https://dashboards.cbs.nl

Dutch population



17°C

○ Basemap
● Daytime Population
○ Incoming Population
○ Outgoing Population

Population density
- Very Quiet
- Quiet
- Normal
- Busy
- Very Busy

Leaflet | © OpenStreetMap contributors © CARTO

Friday 2018-06-01 15:00

Municipality

Amsterdam

Dissemination / Open Data

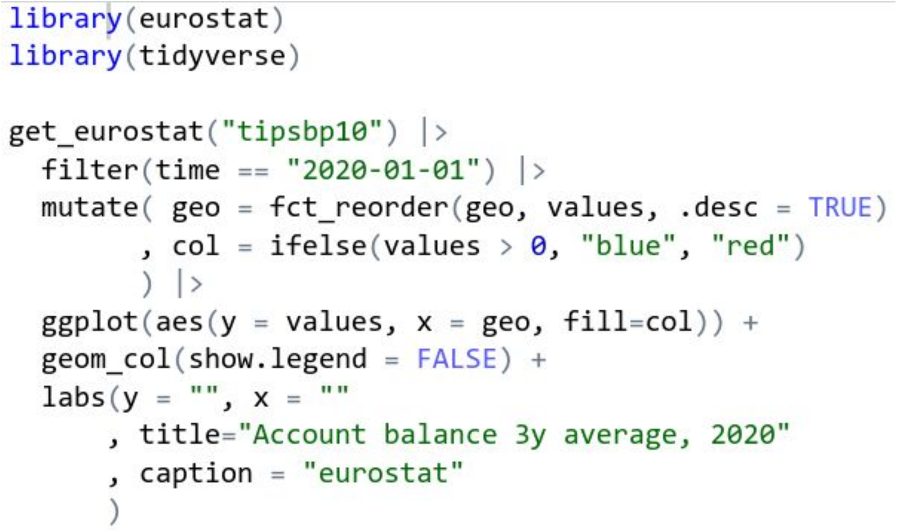
Dissemination

Most National Statistical Agencies have an output database with an open data API

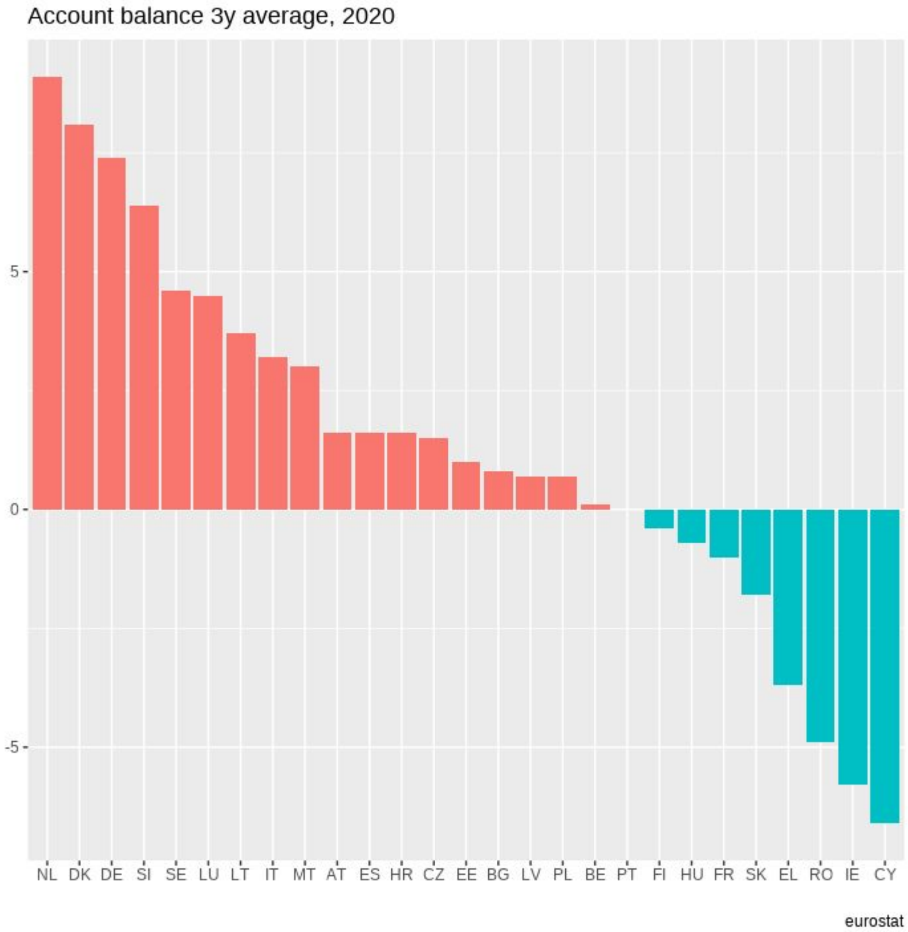
e.g.

- Eurostat: eurostat
- Census Bureau: tidycensus
- Worldbank: wbstats, WDI
- …
- Statistics Netherlands: cbsodataR

Eurostat:



```r
library(eurostat)
library(tidyverse)

get_eurostat("tipsbp10") |>
  filter(time == "2020-01-01") |>
  mutate( geo = fct_reorder(geo, values, .desc = TRUE)
        , col = ifelse(values > 0, "blue", "red")
        ) |>
  ggplot(aes(y = values, x = geo, fill=col)) +
  geom_col(show.legend = FALSE) +
  labs(y = "", x = ""
      , title="Account balance 3y average, 2020"
      , caption = "eurostat"
      )
```



Account balance 3y average, 2020

eurostat

Creating a R package for your data is a good idea!

Providing an R package to retrieve published data:

- helps the mission of your institute
- unlocks your data / removes barriers
- increases use of official statistics!

Example: cbsodataR

R package to retrieve data from the opendata API of statistics Netherlands:

- table of contents: cbs_get_toc
- search the database: cbs_search
- retrieve metadata: cbs_get_meta
- retrieve data: cbs_get_data

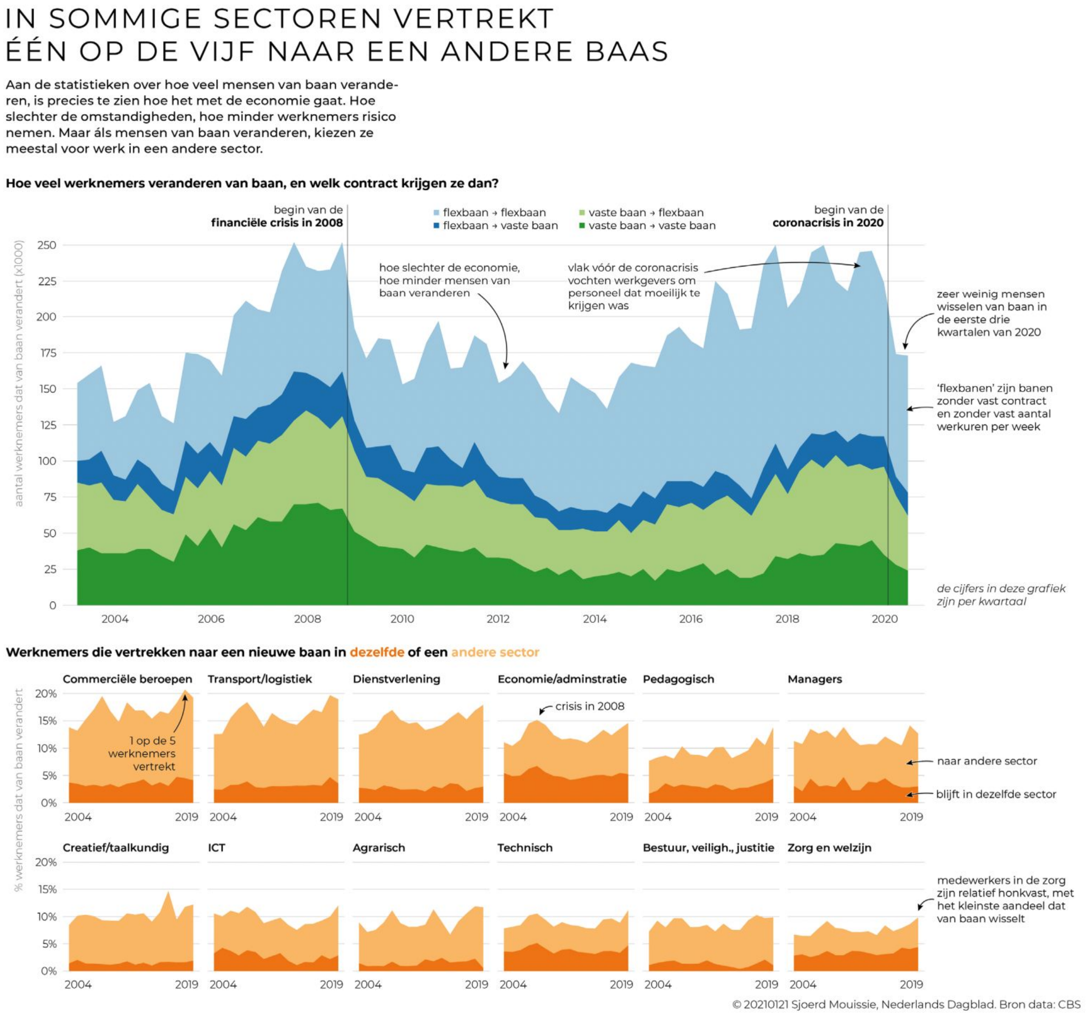
Users of cbsodataR

Government agencies:

- Court of Audit (Rekenkamer) (on twitter “package of the month”).
- Netherlands Environmental Assessment Agency (PBL)
- Netherlands Bureau for Economic Policy Analysis.(CBP)
- Public Health Department (RIVM)
- Government departments
- Municipalities

But also journalists:

data retrieved and graphics prepared with R.
(Job mobility,
"In some sectors,
one out of five changes employer")



IN SOMMIGE SECTOREN VERTREKT ÉÉN OP DE VIJF NAAR EEN ANDERE BAAS

Aan de statistieken over hoe veel mensen van baan veranderen, is precies te zien hoe het met de economie gaat. Hoe slechter de omstandigheden, hoe minder werknemers risico nemen. Maar áls mensen van baan veranderen, kiezen ze meestal voor werk in een andere sector.

Hoe veel werknemers veranderen van baan, en welk contract krijgen ze dan?

begin van de **financiële crisis in 2008**

begin van de **coronacrisis in 2020**

hoe slechter de economie, hoe minder mensen van baan veranderen

vlak vóór de coronacrisis vochten werkgevers om personeel dat moeilijk te krijgen was

zeer weinig mensen wisselen van baan in de eerste drie kwartalen van 2020

'flexbanen' zijn banen zonder vast contract en zonder vast aantal werkuren per week

de cijfers in deze grafiek zijn per kwartaal

Werknemers die vertrekken naar een nieuwe baan in dezelfde of een andere sector

1 op de 5 werknemers vertrekt

crisis in 2008

naar andere sector

blijft in dezelfde sector

medewerkers in de zorg zijn relatief honkvast, met het kleinste aandeel dat van baan wisselt

© 20210121 Sjoerd Mouissie, Nederlands Dagblad. Bron data: CBS

Future of R in Official Statistics

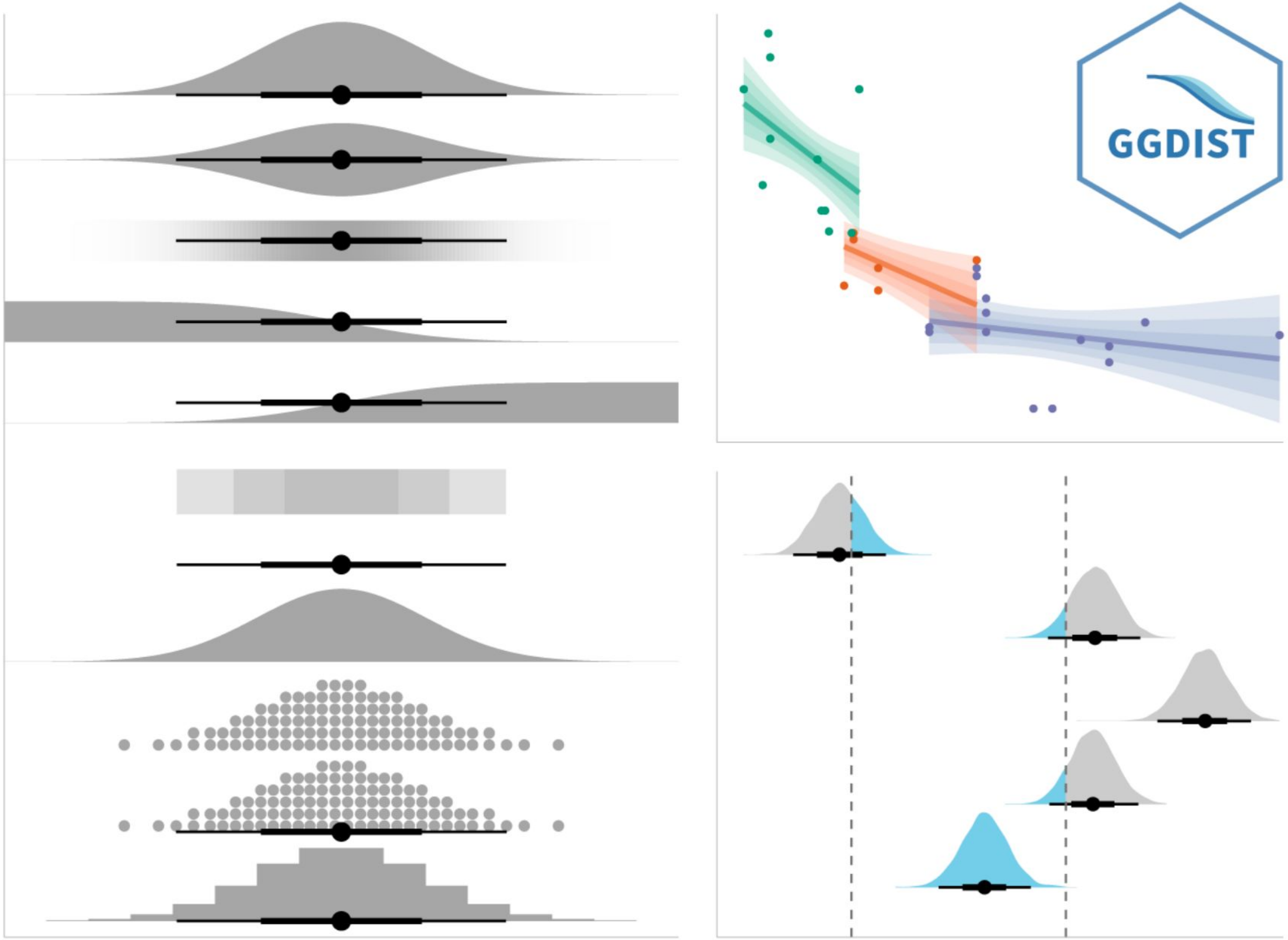
Current state

- R is flourishing in statistical offices
- It is a main language for data science (outside the offices)
- many uses of R (can) support the statistical process of a statistical office.

Directions: Go and explore

To further improve use R to explore:

- data-patterns, outliers etc.
- new **statistical** methods: e.g. rstan
- new **processing** methods: e.g. duckdb / fst
- new **visualisation** methods: e.g ggdist

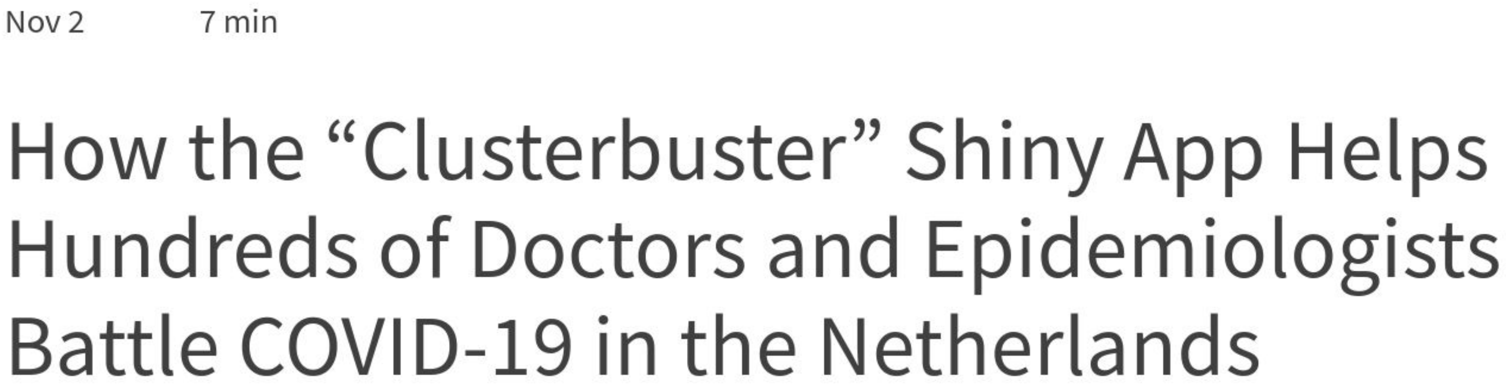

GGDIST

Dashboard / report

Use R to

- communicate, report and popularize statistics from our offices
- e.g. ggplot is used by BBC (bbplot), Financial Times (FT), New York Times
- shiny dashboard are used in the Covid Crisis

Public Health Shiny Dashboard



Nov 2 7 min

How the “Clusterbuster” Shiny App Helps Hundreds of Doctors and Epidemiologists Battle COVID-19 in the Netherlands

Directions: production systems

- Many production systems are now build with R

- Improve their robustness:
  - package dependency, e.g. R package: `renv`
  - data dependency: e.g R package `targets`
  - R-version dependency: e.g. renv or Docker

Directions: other languages

R’s position is good

- R is powerful (homo-iconic)
- has the most official statistics specific packages (> 70%)

Embrace and learn from other languages.

- e.g. Python / Julia / SQL and use the R tools to connect with them
- they can live side-by-side and have their uses

Thanks for your attention!

Questions?

e.dejonge@cbs.nl

@edwindjonge