

Piotr Podlewski

- Member of a multinational programming team working with Real World Data
- Tools: SAS, SQL, R
- 2 years SAS and 4 years R experience



Is R language reliable and efficient tool for programming SAS datasets or just art for art's sake?

Agenda

- Introduction to R
- Data transformation in R
- Do it smart with dplyr!
- Performance
- Does R stand for reliability?
- Recommended reading
- Summary and questions

Introduction to

- The open source software environment and language
- Wide variety of packages
- The development is community driven
- Case sensitivity and no semi-colons
-  Studio is the most popular IDE

Data transformation in R

- R can be used for data modification
- Data frames and other structures
- The tidyverse ecosystem
- Pipelining %>%



```
result <- (v1-v2) ^2 %>% sum() %>% sqrt()
```

Do it smart with dplyr!

Six basic verbs:

arrange, filter, group_by, mutate, select and summarise

```
taadmin <- read_sas("H:/r/taadmin.sas7bdt")
```

```
taadmin2 <- taadmin %>%
```

```
filter(complete.cases(taadmin[["DOSECUN"]])) %>%
```

```
select(INV, PT, DCMDATE, DOSECUN, DOSETL)
```

Do it smart with dplyr!

```
taadminagg <- group_by(taadmin, PT) %>%  
  summarise(sum = sum(DOSETL)) %>% filter(sum>0)
```

```
taadmin <- left_join (taadmin, taadmingr,  
  by = c("PT1" = "PT2", "DT1" = "DT2"))
```

```
taadmin <- taadmin %>% group_by(PT) %>%  
  summarise(avg_d=mean(DOSECUN)) %>%  
  arrange(desc(avg_d))
```

Do it smart with dplyr!

	STUDY	DCMNAME	DCMSUBNM	SUBSETSN	DOCNUM	INVSITE	INV	PT	ACCESSTS	LOGINTS	LSTCHGTS	LOCKFLAG
1	SD12345678	TA_ADMIN	TAA	1	S273034033	511	GRIDELL_C	1561101	2004-08-04 01:47:56	2004-08-03 07:16:23	2005-05-16 07:28:10	Y
2	SD12345678	TA_ADMIN	TAA	1	S296192633	513	MATTIOL_R	1561301	2004-09-07 02:03:27	2004-09-02 03:27:12	2005-04-27 06:10:01	Y
3	SD12345678	TA_ADMIN	TAA	1	S299169233	511	GRIDELL_C	1561101	2004-09-21 01:32:59	2004-09-20 00:30:30	2006-01-04 06:22:40	Y
4	SD12345678	TA_ADMIN	TAA	1	S299170233	511	GRIDELL_C	1561101	2004-09-21 02:53:33	2004-09-20 00:57:04	2006-01-16 10:43:19	Y
5	SD12345678	TA_ADMIN	TAA	1	S299276233	511	GRIDELL_C	1561102	2004-09-21 03:27:53	2004-09-21 01:43:13	2005-04-27 07:00:15	Y
6	SD12345678	TA_ADMIN	TAA	1	S299284033	513	MATTIOL_R	1561301	2004-09-21 05:49:54	2004-09-21 03:25:39	2005-04-27 06:19:19	Y
7	SD12345678	TA_ADMIN	TAA	1	S299287733	513	MATTIOL_R	1561301	2004-09-21 06:33:20	2004-09-21 04:40:45	2005-04-27 06:17:25	Y
8	SD12345678	TA_ADMIN	TAA	1	S299296233	506	DUQUE_A	1560601	2004-09-30 04:01:20	2004-09-21 07:31:11	2005-05-16 07:26:10	Y
9	SD12345678	TA_ADMIN	TAA	1	S303007133	507	CASTELL_J	1560701	2004-10-08 03:53:14	2004-10-08 03:03:37	2005-05-16 07:09:25	Y
10	SD12345678	TA_ADMIN	TAA	1	S308700333	511	GRIDELL_C	1561102	2004-11-12 07:01:59	2004-11-12 01:51:25	2005-04-27 07:01:23	Y

```
taadmin <- select (taadmin, STUDY:SUBSETSN,  
  starts_with("L"))
```

Performance

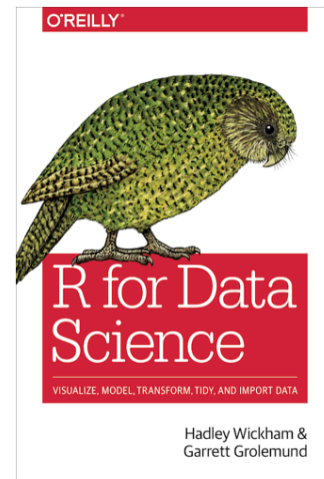
- R loads all data into memory
- Reducing the size of the file before loading
- Improving performance methods- vectorizing and parallel computing
- Improving R's performance by the community

Does stand for reliability?

- Absolutely no warranty
- CRAN, GitHub, RForge or BioConductor?
- Validation coverage is entirely up to the company
- Validated versions of R
- Own repository

Recommended reading

- PHUSE conference archive
- R and RStudio regulatory compliance
- r-bloggers.com
- H. Wickham's & G. Grolemund's books



Summary

- R can be used for programming datasets
- The development of R is community driven
- The dplyr package is a great tool for data modification
- When used wisely, R is efficient and reliable

Do you have any questions?



Paper DH14 „Is R language reliable and efficient tool for programming SAS datasets or just art for art’s sake?”

Thank you for your attention!
piotr.podlewski@quanticate.com

Back-up

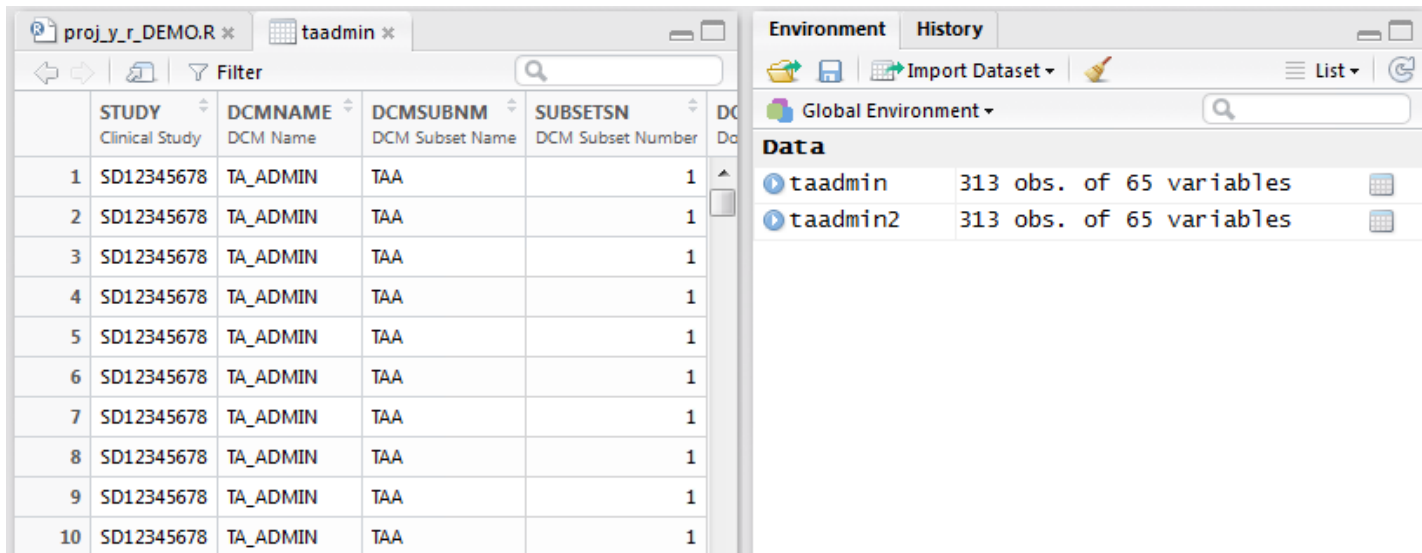
“R: Regulatory Compliance and Validation Issues. A Guidance Document for the Use of R in Regulated Clinical Trial Environments”, 2018

<https://www.r-project.org/doc/R-FDA.pdf>

“RStudio: Regulatory Compliance and Validation Issues. A Guidance Document for the Use of RStudio Professional Products in Regulated Clinical Trial Environments”, 2019

https://rstudio.com/wp-content/uploads/2019/06/rstudio_compliance_validation.pdf

Back-up



The screenshot shows the RStudio interface. The main editor displays a data frame with 10 rows and 5 columns. The columns are labeled STUDY, DCMNAME, DCMSUBNM, SUBSETSN, and DCM. The rows contain data for various clinical studies. The Environment pane on the right shows the Global Environment with two datasets: taadmin and taadmin2, both containing 313 observations and 65 variables.

	STUDY Clinical Study	DCMNAME DCM Name	DCMSUBNM DCM Subset Name	SUBSETSN DCM Subset Number	DCM
1	SD12345678	TA_ADMIN	TAA	1	
2	SD12345678	TA_ADMIN	TAA	1	
3	SD12345678	TA_ADMIN	TAA	1	
4	SD12345678	TA_ADMIN	TAA	1	
5	SD12345678	TA_ADMIN	TAA	1	
6	SD12345678	TA_ADMIN	TAA	1	
7	SD12345678	TA_ADMIN	TAA	1	
8	SD12345678	TA_ADMIN	TAA	1	
9	SD12345678	TA_ADMIN	TAA	1	
10	SD12345678	TA_ADMIN	TAA	1	

Environment History

Global Environment

Data

- taadmin 313 obs. of 65 variables
- taadmin2 313 obs. of 65 variables