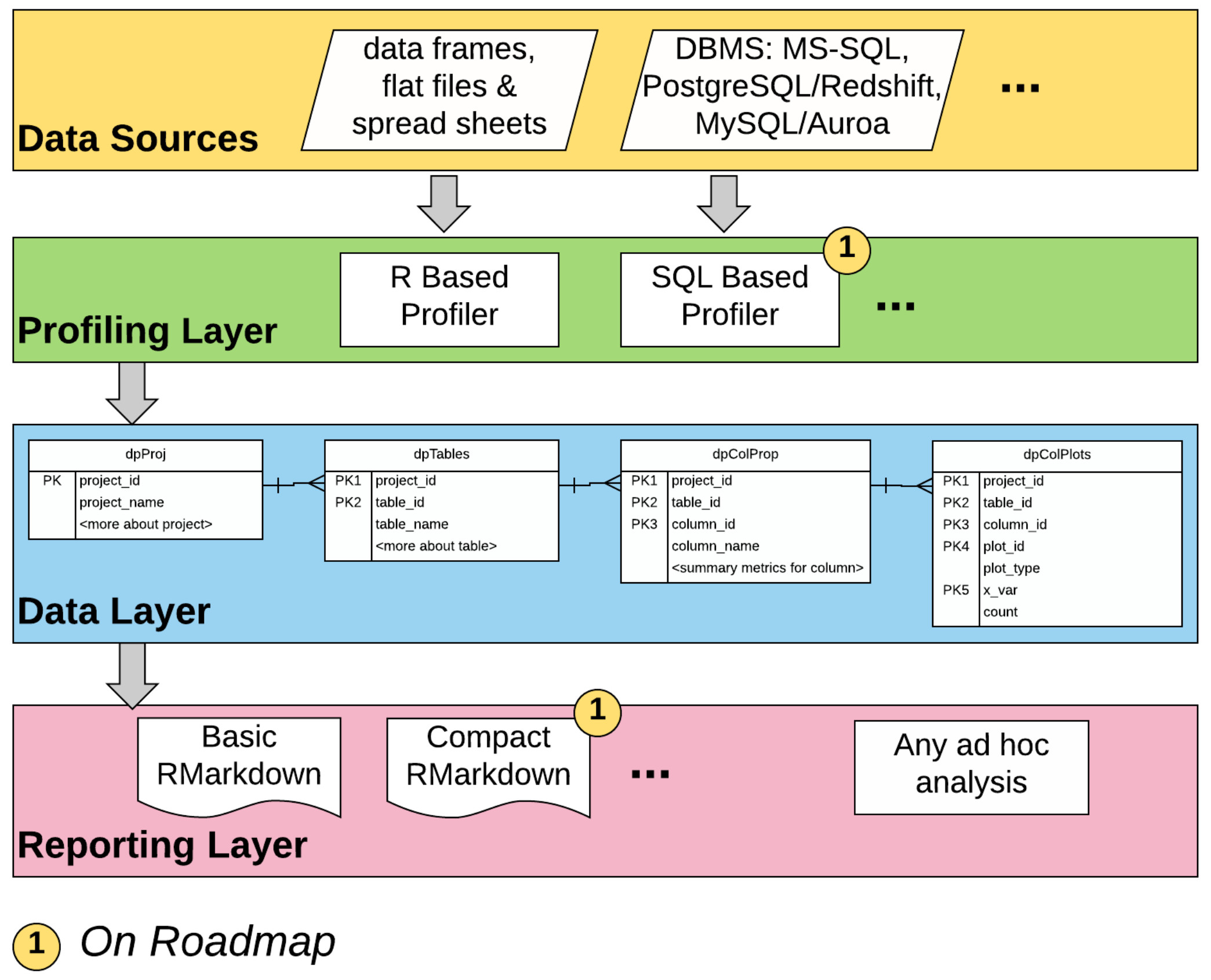
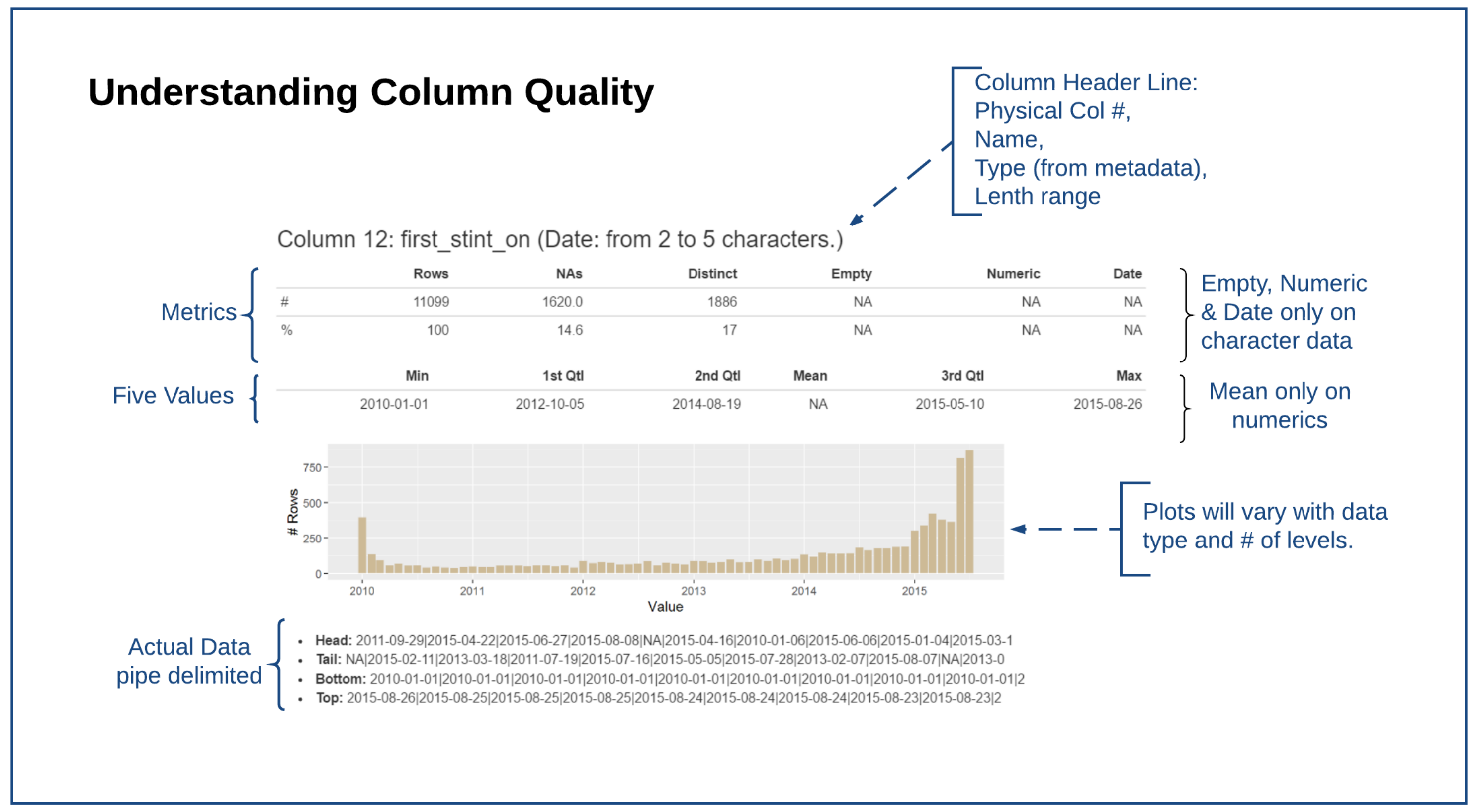
dProf – A Data Quality Profiler by Jim Porzak, DS4CI.org

The dProf package is a total re-write of data quality profiling code I talked about at useR! 2006. That work was inspired by Jack Olson's *Data Quality, The Accuracy Dimension* as is this version.

The main differences are this version makes use of modern R tools like dplyr, readr, and Rmarkdown. More importantly the profiling and presentation functions are de-coupled gives us the flexibility to optimize for different sources and reporting needs:



The Basic RMarkdown module outputs this block of information for each column in the project:



Understanding Column Quality

Column Header Line:
Physical Col #,
Name,
Type (from metadata),
Lenth range

Column 12: first_stint_on (Date: from 2 to 5 characters.)

Metrics

	Rows	NAs	Distinct	Empty	Numeric	Date
#	11099	1620.0	1886	NA	NA	NA
%	100	14.6	17	NA	NA	NA

Empty, Numeric & Date only on character data

Five Values

Min	1st Qtl	2nd Qtl	Mean	3rd Qtl	Max
2010-01-01	2012-10-05	2014-08-19	NA	2015-05-10	2015-08-26

Mean only on numerics

Plots will vary with data type and # of levels.

Actual Data pipe delimited

- **Head:** 2011-09-29|2015-04-22|2015-06-27|2015-08-08|NA|2015-04-16|2010-01-06|2015-06-06|2015-01-04|2015-03-1
- **Tail:** NA|2015-02-11|2013-03-18|2011-07-19|2015-07-16|2015-05-05|2015-07-28|2013-02-07|2015-08-07|NA|2013-0
- **Bottom:** 2010-01-01|2010-01-01|2010-01-01|2010-01-01|2010-01-01|2010-01-01|2010-01-01|2010-01-01|2010-01-01|2
- **Top:** 2015-08-26|2015-08-25|2015-08-25|2015-08-25|2015-08-24|2015-08-24|2015-08-24|2015-08-23|2015-08-23|2

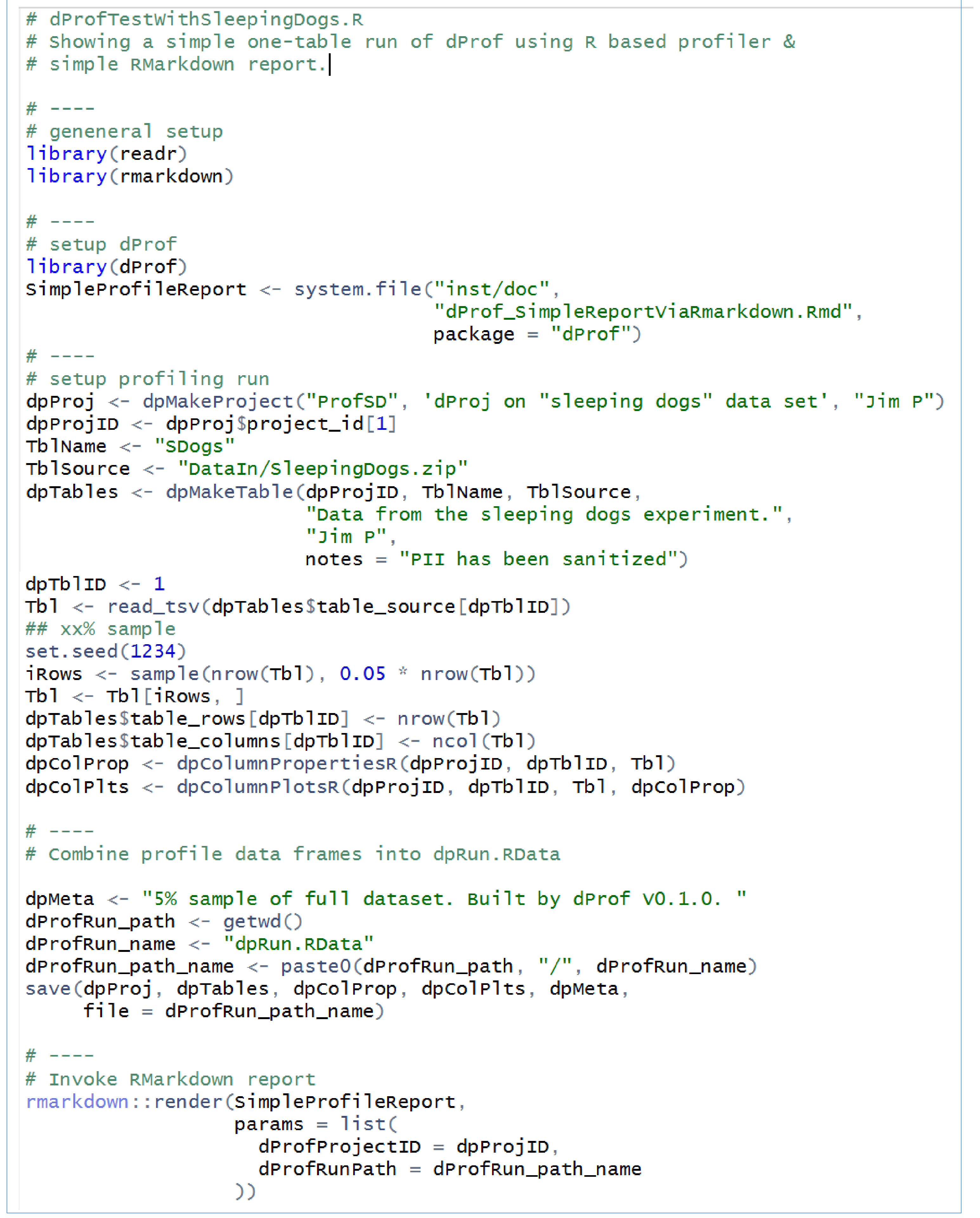
Links:

- github.com/ds4ci/dProf
- Archive: ds4ci.org/archives/
  - ^F "data profiling with R" for original 2006 version
- Email: Jim@DS4CI.org

dProf Workflow

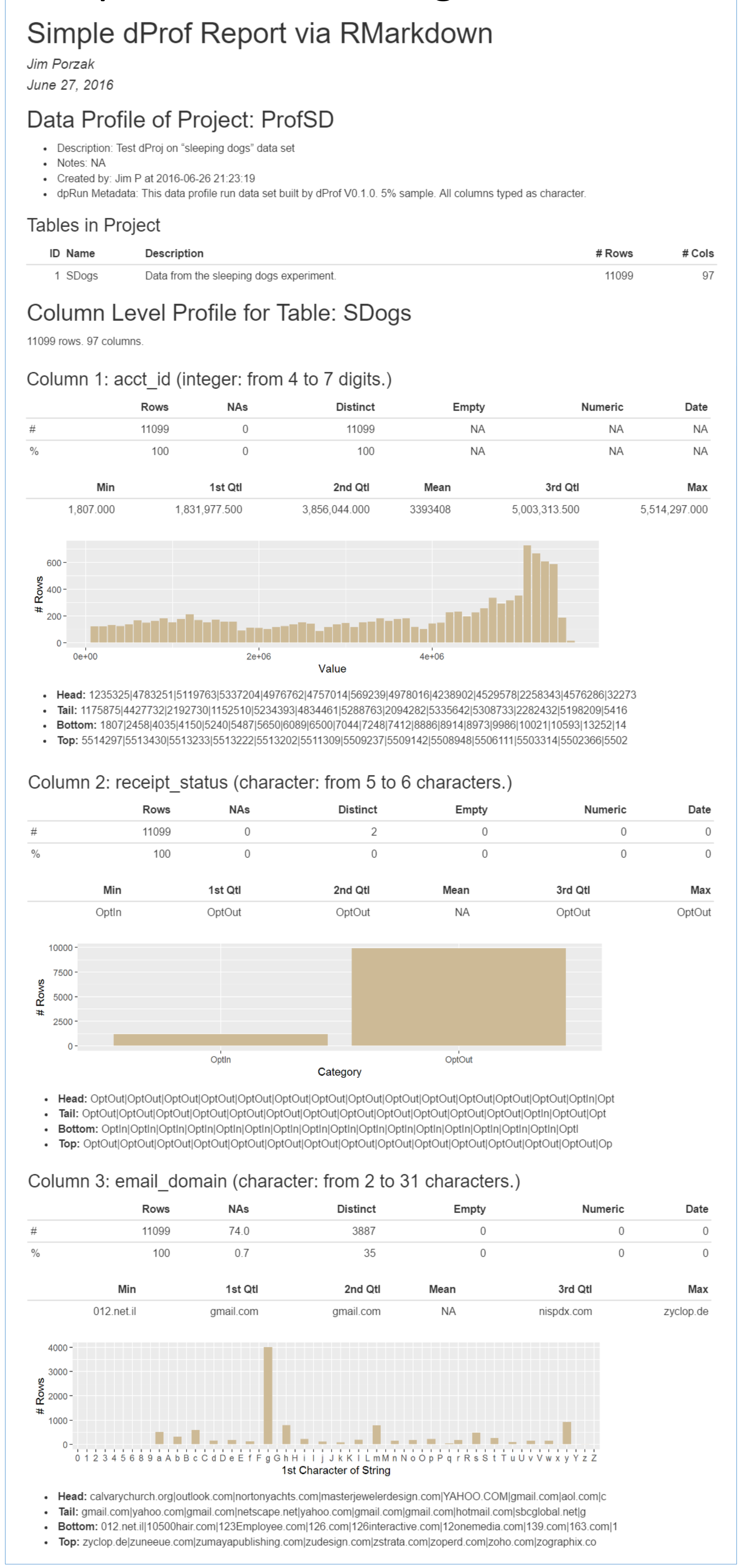
1. Set-up dpProj data frame with `dpMakeProject()`.
2. Set-up dpTables data frame with `dpMakeTable()`.
3. If using the R based profiler, read or load the source files.
4. Profile columns into dpColProp with `dpColumnPropertiesXXX()`, where XXX is "R" or "SQL".
5. Generate plot data into dpColPlts with `dpColumnPlotsXXX()`, where XXX is "R" or "SQL".
6. Build .Rdata file from dpProj, dpTables, dpColProp, dpColPlts, and dpMeta – a character string.
7. Invoke the reporting module of choice – right now that is dProf_SimpleReportViaRmarkdown.Rmd

Sample Run – R Script



```r
# dProfTestWithSleepingDogs.R
# Showing a simple one-table run of dProf using R based profiler &
# simple RMarkdown report.

# ----
# geneneral setup
library(readr)
library(rmarkdown)

# ----
# setup dProf
library(dProf)
SimpleProfileReport <- system.file("inst/doc",
                                   "dProf_SimpleReportViaRmarkdown.Rmd",
                                   package = "dProf")
# ----
# setup profiling run
dpProj <- dpMakeProject("ProfSD", 'dProj on "sleeping dogs" data set', "Jim P")
dpProjID <- dpProj$project_id[1]
TblName <- "SDogs"
TblSource <- "DataIn/SleepingDogs.zip"
dpTables <- dpMakeTable(dpProjID, TblName, TblSource,
                        "Data from the sleeping dogs experiment.",
                        "Jim P",
                        notes = "PII has been sanitized")
dpTblID <- 1
Tbl <- read_tsv(dpTables$table_source[dpTblID])
## xx% sample
set.seed(1234)
iRows <- sample(nrow(Tbl), 0.05 * nrow(Tbl))
Tbl <- Tbl[iRows, ]
dpTables$table_rows[dpTblID] <- nrow(Tbl)
dpTables$table_columns[dpTblID] <- ncol(Tbl)
dpColProp <- dpColumnPropertiesR(dpProjID, dpTblID, Tbl)
dpColPlts <- dpColumnPlotsR(dpProjID, dpTblID, Tbl, dpColProp)

# ----
# Combine profile data frames into dpRun.RData

dpMeta <- "5% sample of full dataset. Built by dProf V0.1.0. "
dProfRun_path <- getwd()
dProfRun_name <- "dpRun.RData"
dProfRun_path_name <- paste0(dProfRun_path, "/", dProfRun_name)
save(dpProj, dpTables, dpColProp, dpColPlts, dpMeta,
     file = dProfRun_path_name)

# ----
# Invoke RMarkdown report
rmarkdown::render(SimpleProfileReport,
                  params = list(
                    dProfProjectID = dpProjID,
                    dProfRunPath = dProfRun_path_name
                  ))
```

Sample Run – Showing 1st 3 Columns



Simple dProf Report via RMarkdown

Jim Porzak

June 27, 2016

Data Profile of Project: ProfSD

- Description: Test dProj on "sleeping dogs" data set
- Notes: NA
- Created by: Jim P at 2016-06-26 21:23:19
- dpRun Metadata: This data profile run data set built by dProf V0.1.0. 5% sample. All columns typed as character.

Tables in Project

ID	Name	Description	# Rows	# Cols
1	SDogs	Data from the sleeping dogs experiment.	11099	97

Column Level Profile for Table: SDogs

11099 rows. 97 columns.

Column 1: acct_id (integer: from 4 to 7 digits.)

	Rows	NAs	Distinct	Empty	Numeric	Date
#	11099	0	11099	NA	NA	NA
%	100	0	100	NA	NA	NA

Min	1st Qtl	2nd Qtl	Mean	3rd Qtl	Max
1,807.000	1,831,977.500	3,856,044.000	3393408	5,003,313.500	5,514,297.000

- **Head:** 1235325|4783251|5119763|5337204|4976762|4757014|569239|4978016|4238902|4529578|2258343|4576286|32273
- **Tail:** 1175875|4427732|2192730|1152510|5234393|4834461|5288763|2094282|5335642|5308733|2282432|5198209|5416
- **Bottom:** 1807|2458|4035|4150|5240|5487|5650|6089|6500|7044|7248|7412|8886|8914|8973|9986|10021|10593|13252|14
- **Top:** 5514297|5513430|5513233|5513222|5513202|5511309|5509237|5509142|5508948|5506111|5503314|5502366|5502

Column 2: receipt_status (character: from 5 to 6 characters.)

	Rows	NAs	Distinct	Empty	Numeric	Date
#	11099	0	2	0	0	0
%	100	0	0	0	0	0

Min	1st Qtl	2nd Qtl	Mean	3rd Qtl	Max
OptIn	OptOut	OptOut	NA	OptOut	OptOut

- **Head:** OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptIn|Opt
- **Tail:** OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptIn|OptOut|Opt
- **Bottom:** OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptIn|OptI
- **Top:** OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|OptOut|Op

Column 3: email_domain (character: from 2 to 31 characters.)

	Rows	NAs	Distinct	Empty	Numeric	Date
#	11099	74.0	3887	0	0	0
%	100	0.7	35	0	0	0

Min	1st Qtl	2nd Qtl	Mean	3rd Qtl	Max
012.net.il	gmail.com	gmail.com	NA	nispdx.com	zyclop.de

# Rows
4000
3000
2000
1000
0
0 1 2 3 4 5 6 8 9 a A b B c C d D e E f F g G h H i I j J k K l L m M n N o O p P q r R s S t T u U v V w x y Y z Z
1st Character of String

- **Head:** calvarychurch.org|outlook.com|nortonyachts.com|masterjewelerdesign.com|YAHOO.COM|gmail.com|aol.com|c
- **Tail:** gmail.com|yahoo.com|gmail.com|netscape.net|yahoo.com|gmail.com|gmail.com|hotmail.com|sbcglobal.net|g
- **Bottom:** 012.net.il|10500hair.com|123Employee.com|126.com|126interactive.com|12onemedia.com|139.com|163.com|1
- **Top:** zyclop.de|zuneeue.com|zumayapublishing.com|zudesign.com|zstrata.com|zoperd.com|zoho.com|zographix.co