Web Codebook: Interactive Data Set Summaries

Jeremy Wildfire, Ryan Bailey, Spencer Childress and Rebecca Krouse

Rho, Inc.

What is a codebook?

Codebook – A document that describes and summarizes the variables in a dataset.

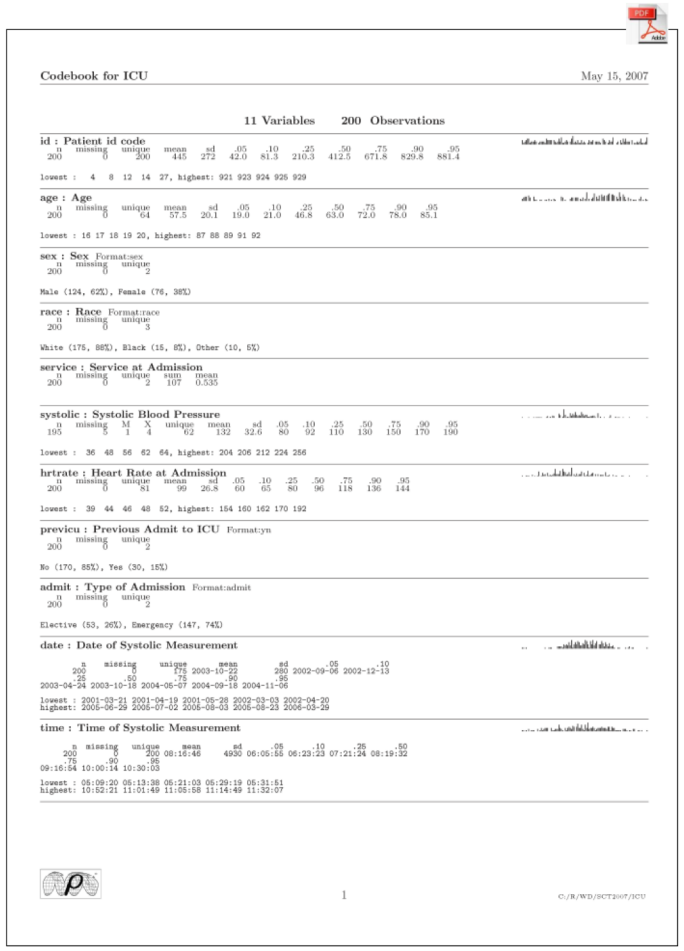
Codebooks contain information about:

- Variable distributions
- Summary Statistics
- Basic Listing
- Missing data summary
- Variable metadata – variable format, labels, etc.

Existing Codebook Tools

hmisc::summarize (R) by Frank Harrell

Link: R Package



Codebook for ICU — May 15, 2007

11 Variables 200 Observations

id : Patient id code

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	200	445	272	42.0	81.3	210.3	412.5	671.8	829.8	881.4

lowest : 4 8 12 14 27, highest: 921 923 924 925 929

age : Age

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	64	57.5	20.1	19.0	21.0	46.8	63.0	72.0	78.0	85.1

lowest : 16 17 18 19 20, highest: 87 88 89 91 92

sex : Sex Format:sex

n	missing	unique
200	0	2

Male (124, 62%), Female (76, 38%)

race : Race Format:race

n	missing	unique
200	0	3

White (175, 88%), Black (15, 8%), Other (10, 5%)

service : Service at Admission

n	missing	unique	sum	mean
200	0	2	107	0.535

systolic : Systolic Blood Pressure

n	missing	M	X	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
195	5	1	4	62	132	32.6	80	92	110	130	150	170	190

lowest : 36 48 56 62 64, highest: 204 206 212 224 256

hrtrate : Heart Rate at Admission

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	81	99	26.8	60	65	80	96	118	136	144

lowest : 39 44 46 48 52, highest: 154 160 162 170 192

previcu : Previous Admit to ICU Format:yn

n	missing	unique
200	0	2

No (170, 85%), Yes (30, 15%)

admit : Type of Admission Format:admit

n	missing	unique
200	0	2

Elective (53, 26%), Emergency (147, 74%)

date : Date of Systolic Measurement

n	missing	unique	mean	sd	.05	.10
200	0	175	2003-10-22	280	2002-09-06	2002-12-13
.25	.50	.75	.90	.95		
2003-04-24	2003-10-18	2004-05-07	2004-09-18	2004-11-06		

lowest : 2001-03-21 2001-04-19 2001-05-28 2002-03-03 2002-04-20
highest: 2005-06-29 2005-07-02 2005-08-03 2005-08-23 2006-03-29

time : Time of Systolic Measurement

n	missing	unique	mean	sd	.05	.10	.25	.50
200	0	200	08:16:46	4930	06:05:55	06:23:23	07:21:24	08:19:32
.75	.90	.95						
09:16:54	10:00:14	10:30:03						

lowest : 05:09:20 05:13:38 05:21:03 05:29:19 05:31:51
highest: 10:52:21 11:01:49 11:05:58 11:14:49 11:32:07

1 C:/R/WD/SCT2007/ICU

Codebook for ICU — May 15, 2007

11 Variables 200 Observations

id : Patient id code

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	200	445	272	42.0	81.3	210.3	412.5	671.8	829.8	881.4

lowest : 4 8 12 14 27, highest: 921 923 924 925 929

age : Age

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	64	57.5	20.1	19.0	21.0	46.8	63.0	72.0	78.0	85.1

lowest : 16 17 18 19 20, highest: 87 88 89 91 92

sex : Sex Format:sex

n	missing	unique
200	0	2

Male (124, 62%), Female (76, 38%)

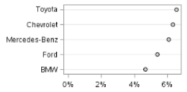
Version shown generated using a SAS Macro from Agustin Calatroni

Existing Codebook Tools

sas-codebook macro by Shane Rosenbalm

Link: GitHub Repo

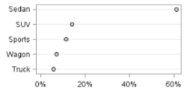
Abbr. codebook for EXAMPLES.CARS (15 vars, 428 obs) Page 1



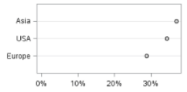
make: -- no label -- {type=char}
n=428, missing=0, unique=38
Frequencies (top 5 only): Toyota (28, 6.5%), Chevrolet (27, 6.3%), Mercedes-Benz (26, 6.1%), Ford (23, 5.4%), BMW (20, 4.7%)

Too many unique values.
No graphic will be produced.

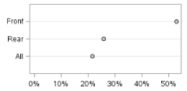
model: -- no label -- {type=char}
n=428, missing=0, unique=425
First 5: 3.5 RL 4dr, 300M 4dr, 325Ci 2dr, 325i 4dr, 325xi 4dr
Last 5: Z4 convertible 3.0i..., Z4 convertible 2.5i..., Wrangler Sahara con..., Viper SRT-10 conver..., Tundra Regular Cab...



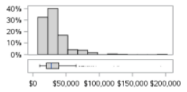
type: -- no label -- {type=char}
n=428, missing=0, unique=6
Frequencies (top 5 only): Sedan (262, 61.2%), SUV (60, 14.0%), Sports (49, 11.4%), Wagon (30, 7.0%), Truck (24, 5.6%)



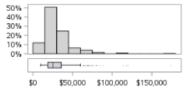
origin: -- no label -- {type=char}
n=428, missing=0, unique=3
Frequencies: Asia (158, 36.9%), USA (147, 34.3%), Europe (123, 28.7%)



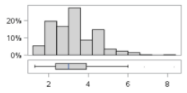
drivetrain: -- no label -- {type=char}
n=428, missing=0, unique=3
Frequencies: Front (226, 52.8%), Rear (110, 25.7%), All (92, 21.5%)



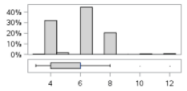
msrp: -- no label -- {type=num, fmt=DOLLAR8.}
n=428, missing=0, unique=410
Mean (StdDev) = $32,775 (19431.72) Min, Max = $10,280, $192,465
Median [Q1, Q3] = $27,635 [$20,330, $39,215]



invoice: -- no label -- {type=num, fmt=DOLLAR8.}
n=428, missing=0, unique=425
Mean (StdDev) = $30,015 (17642.12) Min, Max = $9,875, $173,560
Median [Q1, Q3] = $25,295 [$18,851, $35,733]

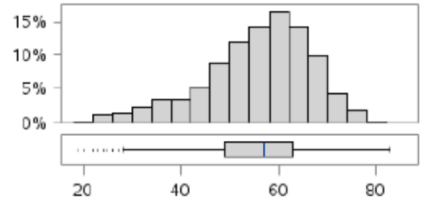


enginesize: Engine Size (L) {type=num}
n=428, missing=0, unique=43
Mean (StdDev) = 3.2 (1.11) Min, Max = 1.3, 8.3
Median [Q1, Q3] = 3 [2.4, 3.9]

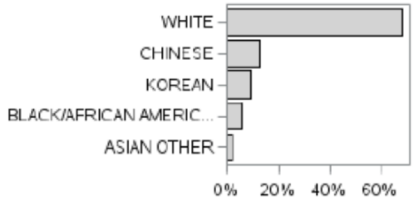


cylinders: -- no label -- {type=num}
n=426, missing=2, unique=7
Mean (StdDev) = 5.8 (1.56) Min, Max = 3, 12
Median [Q1, Q3] = 6 [4, 6]

Shortened text strings are denoted by three dots (...).
H:\GitHub\RhoInc\sas-codebook\ExampleFiles 2016-11-04T14:07:06



age: Drv: Age At Consent Date {type=num}
n=975, missing=0, unique=62
Mean (StdDev) = 55.2 (11.19) Min, Max = 19, 83
Median [Q1, Q3] = 57 [49, 63]



race: Race {type=num, fmt=RACE2F.}
n=975, missing=0, unique=10
Frequencies (top 5 only): 1 = WHITE (670, 68.7%), 19 = CHINESE (125, 12.8%), 21 = KOREAN (91, 9.3%), 2 = BLACK/AFRICAN AMERIC... (53, 5.4%), 91 = ASIAN OTHER (18, 1.8%)

Why are codebooks useful?

Link: [Article]

COMMENT

P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

There is no statistic more maligned than the *P* value. Hundreds of papers and blogposts have been written about what some statisticians deride as 'null hypothesis significance testing' (NHST; see, for example, go.nature.com/pfvgqe). NHST deems whether the results of a data analysis are important on the basis of whether a summary statistic (such as a *P* value) has crossed a threshold. Given the discourse, it is no surprise that some hailed as a victory the banning of NHST methods (and all of statistical inference) in the journal *Basic and Applied Social Psychology* in February[1].

Such a ban will in fact have scant effect on the quality of published science. There are many stages to the design and analysis of a successful study (see 'Data pipeline'). The last of these steps is the calculation of an inferential statistic such as a *P* value, and the application of a 'decision rule' to it (for example, $P < 0.05$). In practice, decisions that are made earlier in data analysis have a much greater impact on results — from experimental design to batch effects, lack of adjustment for confounding factors, or simple measurement error. Arbitrary levels of statistical significance can be achieved by changing the ways in which data are cleaned, summarized or modelled[2].

P values are an easy target: being widely used, they are widely abused. But, in practice, deregulating statistical significance opens the door to even more ways to game statistics — intentionally or unintentionally — to get a result. Replacing *P* values with Bayes factors or another statistic is ultimately about choosing a different trade-off of true positives and false positives. Arguing about the *P* value is like focusing on a single misspelling, rather than on the faulty logic of a sentence.

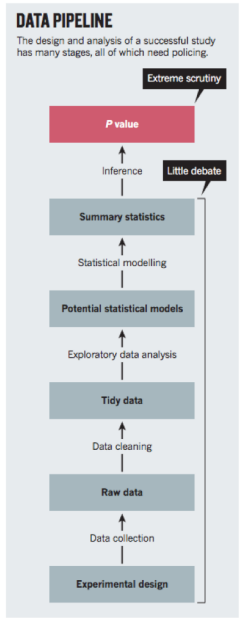
Better education is a start. Just as anyone who does DNA sequencing or remote-sensing has to be trained to use a machine, so too anyone who analyses data must be trained in the relevant software and concepts. Even investigators who supervise data analysis should be required by their funding agencies and institutions to complete training in understanding the outputs and potential problems with an analysis.

There are online courses specifically designed to address this crisis. For example, the Data Science Specialization, offered by Johns Hopkins University in Baltimore, Maryland, and Data Carpentry, can easily be integrated into training and research. It is increasingly possible to learn to use the computing tools relevant to specific disciplines — training in Bioconductor, Galaxy and Python is included in Johns Hopkins' Genomic Data Science Specialization, for instance.

But education is not enough. Data analysis is taught through an apprenticeship model, and different disciplines develop their own analysis subcultures. Decisions are based on cultural conventions in specific communities rather than on empirical evidence. For example, economists call data measured over time 'panel data', to which they frequently apply mixed-effects models. Biomedical scientists refer to the same type of data structure as 'longitudinal data', and often go at it with generalized estimating equations.

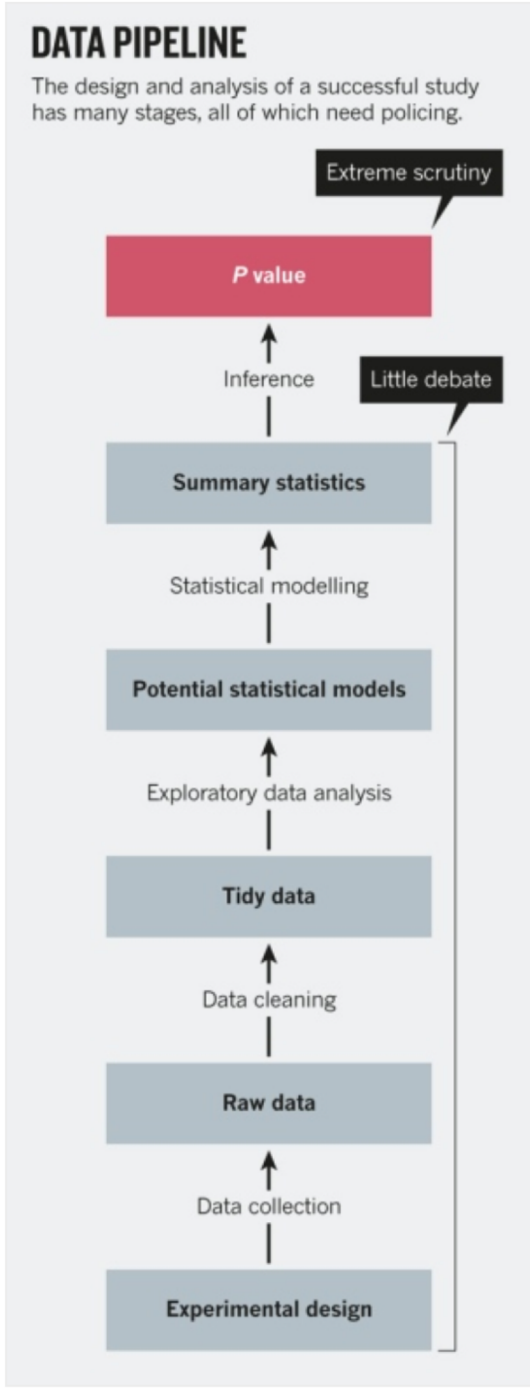
Statistical research largely focuses on mathematical statistics, to the exclusion of the behaviour and processes involved in data analysis. To solve this deeper problem, we must study how people perform data analysis in the real world. What sets them up for success, and what for failure? Controlled experiments have been done in visualization[3] and risk interpretation[4] to evaluate how humans perceive and interact with data and statistics. More recently, we and others have been studying the entire analysis pipeline. We found, for example, that recently trained data analysts do not know how to infer *P* values from plots of data[5], but they can learn to do so with practice.

The ultimate goal is evidence-based data analysis[6]. This is analogous to evidence-based medicine, in which physicians are encouraged to use only treatments for which efficacy has been proved in controlled trials. Statisticians and the people they teach and collaborate with need to stop arguing about *P* values, and prevent the rest of the iceberg from sinking science. ■



Jeffrey T. Leek *and* **Roger D. Peng** *are associate professors of biostatistics at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA.*
e-mail: jleek@jhsph.edu

1. Trafimow, D. & Marks, M. *Basic Appl. Soc. Psych.* **37,** 1–2 (2015).
2. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22,** 1359–1366 (2011).
3. Cleveland, W. S. & McGill, R. *Science* **229,** 828–833 (1985).
4. Kahneman, D. & Tversky, A. *Econometrica* **47,** 263–291 (1979).
5. Fisher, A., Anderson, G. B., Peng, R. & Leek, J. *PeerJ* **2,** e589 (2014).
6. Leek, J. T. & Peng, R. D. *Proc. Natl Acad. Sci. USA* **112,** 1645–1646 (2015).



Codebooks are great for exploring data

Many data errors can be caught with a simple codebook .

Exploratory Data Analysis - Typical Workflow

Request from Investigator: How many people were allergic to roach from our study in 2002?

Me: Uhh, where does that data live? I wonder if they want German Roach? Skin Test or IgE? 30 minutes later, runs this code.

```
> path <- "crazy/path/to/data/from/10/years/ago/sdtm"
> skintests <- import(paste0(path,"/sas/st.sas7bdat"))
> table(skintests$germanroach_skintest)


Positive 
     219 
Negative 
     200 
```

Email to Investigator [after a quick minute trip to my phone's calculator]:
52% (219/419) were allergic to german roach according to skin test.

Exploratory Data Analysis: Improved Workflow

Me: Codebooks were auto-generated every night, let's take a look.

---

germanroach.skintest: German Roach - Wheal Size at least 3.0 mm > Saline Wheal Size Format:yesnofm

n	missing	unique
419	0	2

No (200, 48%), Yes (219, 52%)

---

roachmix.skintest: Roach Mix - Wheal Size at least 3.0 mm > Saline Wheal Size Format:yesnofm

n	missing	unique
419	0	2

No (193, 46%), Yes (226, 54%)

---

roach.ige : IgE Cockroach (IU/mL)

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
417	2	339	3.8	9.37	0.0125	0.0337	0.0727	0.1800	1.5800	12.0617	19.8188

lowest : 0.0000 0.0101 0.0116 0.0128 0.0156
highest: 46.8543 50.9983 51.2511 57.7085 74.4087

roach.igec : Categorized IgE Cockroach Format:igefm

n	missing	unique
396	23	3

0.35-3.5 KU/L (257, 65%), 3.5-20 KU/L (68, 17%)
More than 20 KU/L (71, 18%)

Email to Investigator:
Looks like 52% and 54% were allergic to German cockroach and Roach mix respectively according to skin test. 35% had sIgE above 3.5. I've attached a data set summary with more info including wheal size and exact IgE levels.


But wait ...

Exploratory Data Analysis – The Real Workflow

2 days later ...

Request #2: Great! Can you break this down by Site?

A week passes …

Request #3: What about by gender?

4 days after that …

Request #4: How many are both mouse and cockroach allergic?

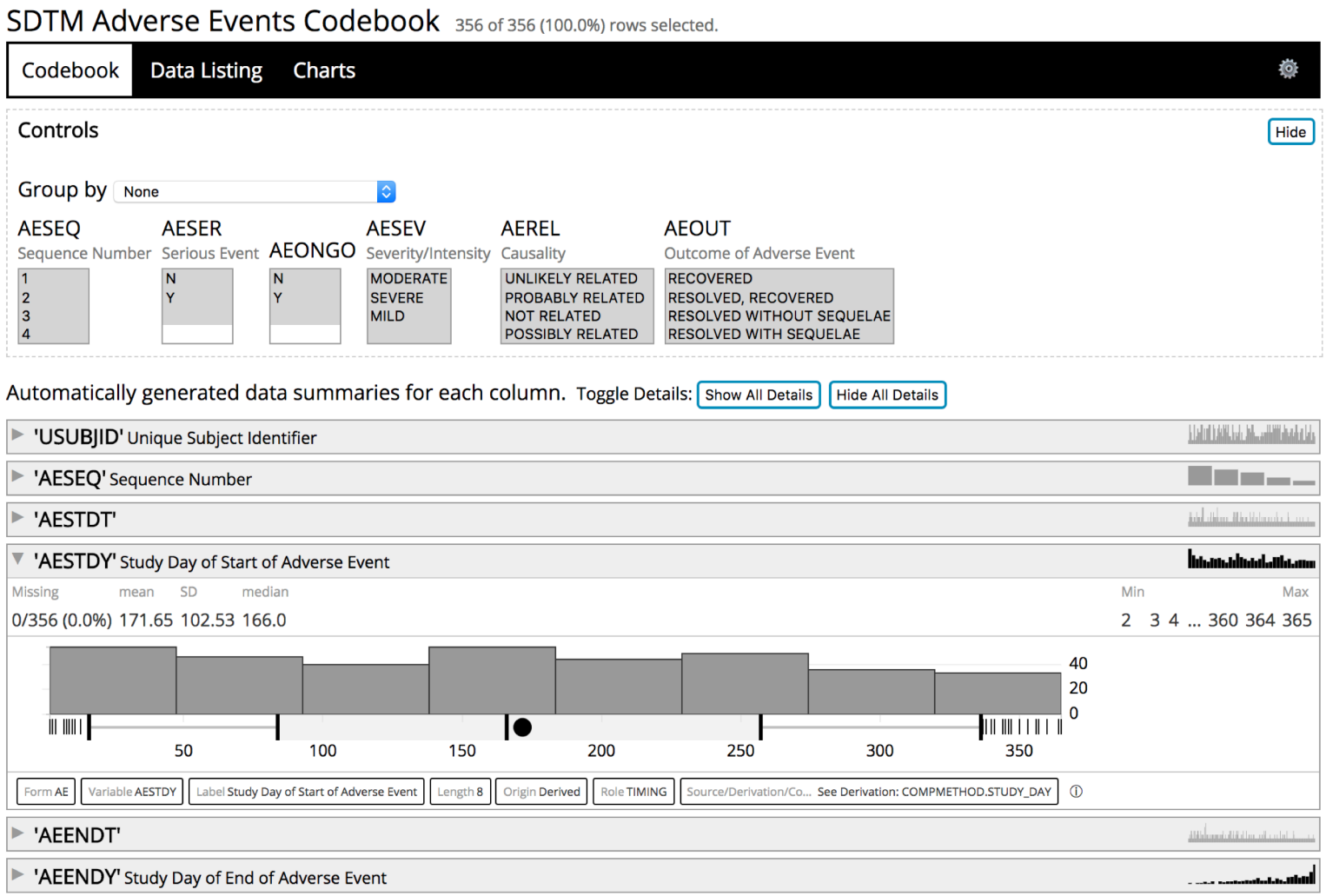
2 months later while you're on vacation …

Request #5: That paper is due tomorrow! Can I have p-values for all of that??

Introducing

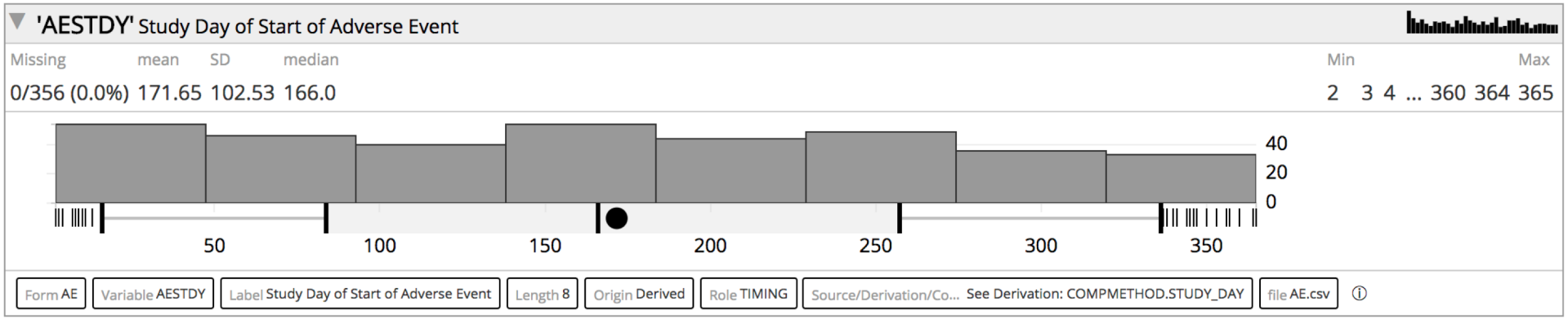
Interactive Web Codebook

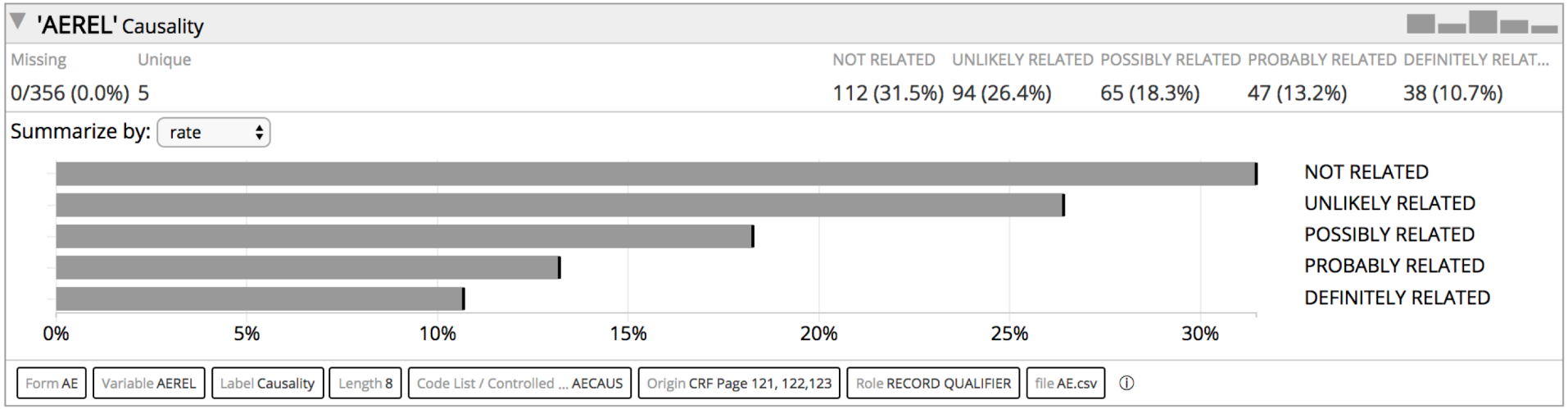
Links: GitHub Repo – Live Example – R Impementation



Summary First …

… Details on Demand





Linked Listing

AESTDY ↓ ✕ Search 356 records displayed

USUBJID	AESEQ	AESTDT	AESTDY	AEENDT	AEENDY	AEDECOD	AESER	AEONGO	AESEV	AEREL	AEOUT
03-013	1	2015-11-06	10	2016-06-03	220	Chronic kidney disease	N	Y	MILD	PROBABLY RELATED	RECOVERED
05-010	1	2015-12-26	10	2016-07-31	228	Azoospermia	N	N	MODERATE	NOT RELATED	RECOVERED
05-022	2	2015-07-24	101	2015-08-29	137	Cognitive disturbance	N	Y	MILD	POSSIBLY RELATED	RECOVERED
05-028	1	2015-06-09	101	2016-02-22	359	CPK increased	N	N	MILD	POSSIBLY RELATED	RESOLVED WITH SEQUELAE
05-025	2	2015-11-20	103	2016-04-19	254	Laryngitis	Y	N	MILD	PROBABLY RELATED	RECOVERED
04-009	4	2015-07-21	103	2015-11-10	215	Serum amylase increased	N	N	MILD	POSSIBLY RELATED	RECOVERED
05-029	2	2016-03-13	103	2016-09-20	294	Laryngeal obstruction	Y	Y	MILD	NOT RELATED	RESOLVED, RECOVERED
01-018	2	2016-03-25	104	2016-12-06	360	Ovarian rupture	Y	N	MILD	POSSIBLY RELATED	RECOVERED
04-004	4	2016-03-01	106	2016-10-09	328	Jejunal hemorrhage	Y	N	MODERATE	DEFINITELY RELATED	RESOLVED WITHOUT SEQUELAE
02-029	2	2016-02-04	106	2016-07-01	254	Middle ear inflammation	N	N	MILD	UNLIKELY RELATED	RECOVERED

Export: CSV

1 2 3 4 5 ... > >>

Explore Multiple Files

Dataset	Description	Class	Structure	Purpose	Keys	Location
TA	Trial Arms	Trial Design	One record per planned Element per Arm	Tabulation	STUDYID, ARMCD, TAETORD	ta.xpt
SUPPLB	Supplemental Qualifiers for LB	Relationship	One record per IDVAR, IDVARVAL, and QNAM value per subject	Tabulation	STUDYID, RDOMAIN, USUBJID, IDVAR, IDVARVAL, QNAM	supplb.xpt
TI	Trial Inclusion/ Exclusion Criteria	Trial Design	One record per I/E criterion	Tabulation	STUDYID, IETESTCD	ti.xpt
TS	Trial Summary	Trial Design	One record per trial summary parameter value	Tabulation	STUDYID, TSPARMCD, TSSEQ	ts.xpt
TV	Trial Visits	Trial Design	One record per planned Visit per Arm	Tabulation	STUDYID, VISITNUM	tv.xpt
DM	Demographics	Special Purpose	One record per subject	Tabulation	STUDYID, USUBJID	dm.xpt
SE	Subject Elements	Special Purpose	One record per actual Element per subject	Tabulation	STUDYID, USUBJID, ETCD	se.xpt
SV	Subject Visits	Special Purpose	One record per actual visit per subject	Tabulation	STUDYID, USUBJID, VISITNUM	sv.xpt
CM	Concomitant Medications	Interventions	One record per recorded medication occurrence or constant-dosing interval per subject	Tabulation	STUDYID, USUBJID, CMTRT, CMSTDTC	cm.xpt
EX	Exposure	Interventions	One record per constant dosing interval per subject	Tabulation	STUDYID, USUBJID, EXTRT, EXSTDTC	ex.xpt
AE	Adverse Events	Events	One record per adverse event per subject	Tabulation	STUDYID, USUBJID, AETERM, AESTDTC, AESEQ	ae.xpt
TE	Trial Elements	Trial Design	One record per planned Element	Tabulation	STUDYID, ETCD	te.xpt
MH	Medical History	Events	One record per medical history event per subject	Tabulation	STUDYID, USUBJID, MHTERM, MHSTDTC	mh.xpt
LB	Laboratory Tests Results	Findings	One record per analyte per planned time point number per time point reference per visit per subject	Tabulation	STUDYID, USUBJID, LBTESTCD, VISITNUM	lb.xpt
QS	Questionnaires	Findings	One record per questionnaire per question per time point per visit per subject	Tabulation	STUDYID, USUBJID, QSTESTCD, VISITNUM	qs.xpt

Real Time Configuration

Column	Label	Group	Filter	Hide
USUBJID	Unique Subject Identifie	☐	☐	☐
AESEQ	Sequence Number	☑	☑	☐
AESTDT		☐	☐	☐
AESTDY	Study Day of Start of Ac	☐	☐	☐
AEENDT		☐	☐	☐
AEENDY	Study Day of End of Adv	☐	☐	☐
AETERM	Reported Term for the A	☐	☐	☐
AEDECOD	Dictionary-Derived Tern	☐	☐	☐
AEBODSYS	Body System or Organ C	☐	☐	☐
AESER	Serious Event	☑	☑	☐
AEONGO		☑	☑	☐
AESEV	Severity/Intensity	☑	☑	☐
AEREL	Causality	☑	☑	☐
AEOUT	Outcome of Adverse Ev	☑	☑	☐

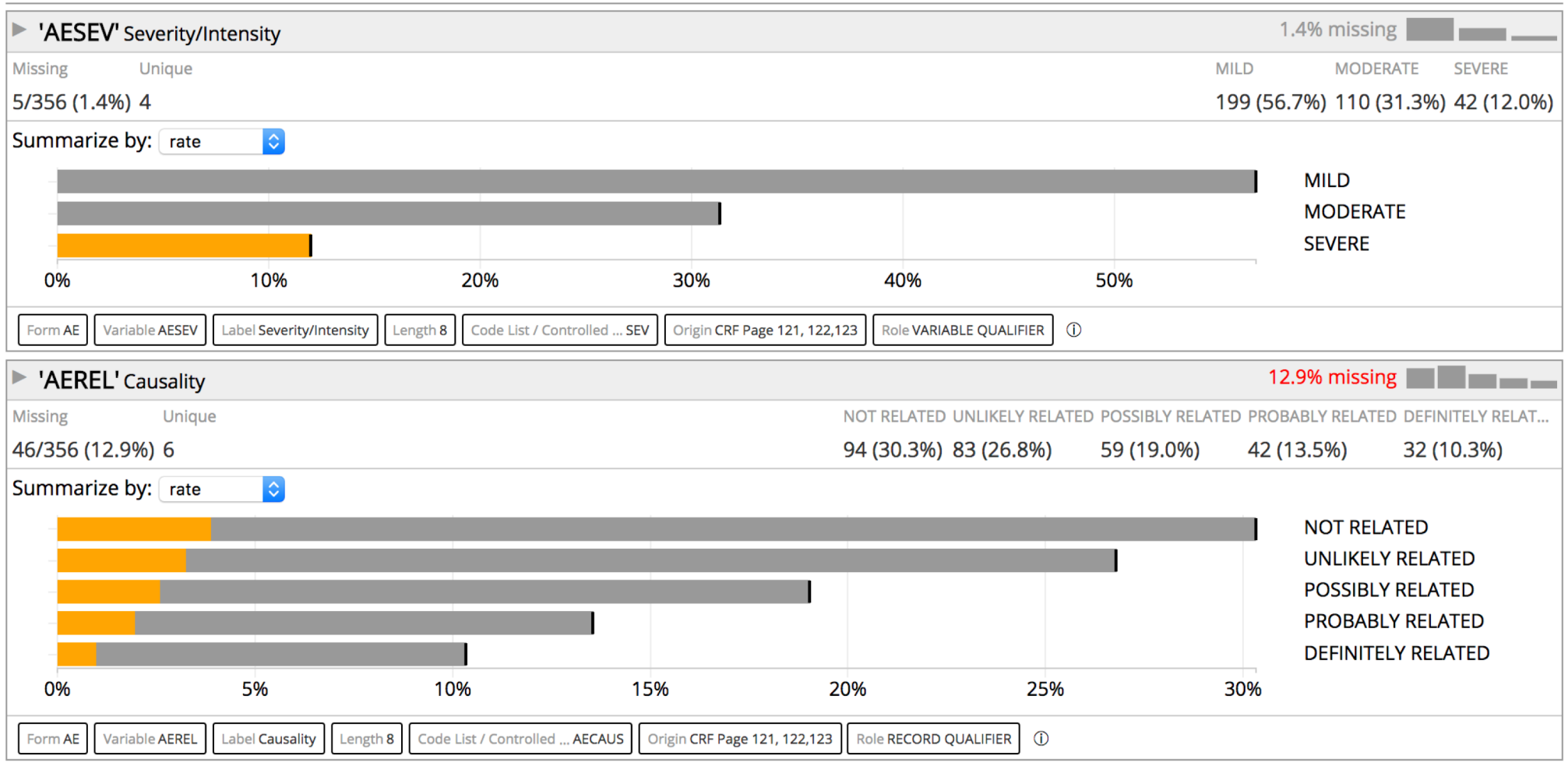
Basic Real-Time Charting

Pick two variables to compare. Filter and group (panel) the chart using the controls above.

x variable: Age y variable: Sex

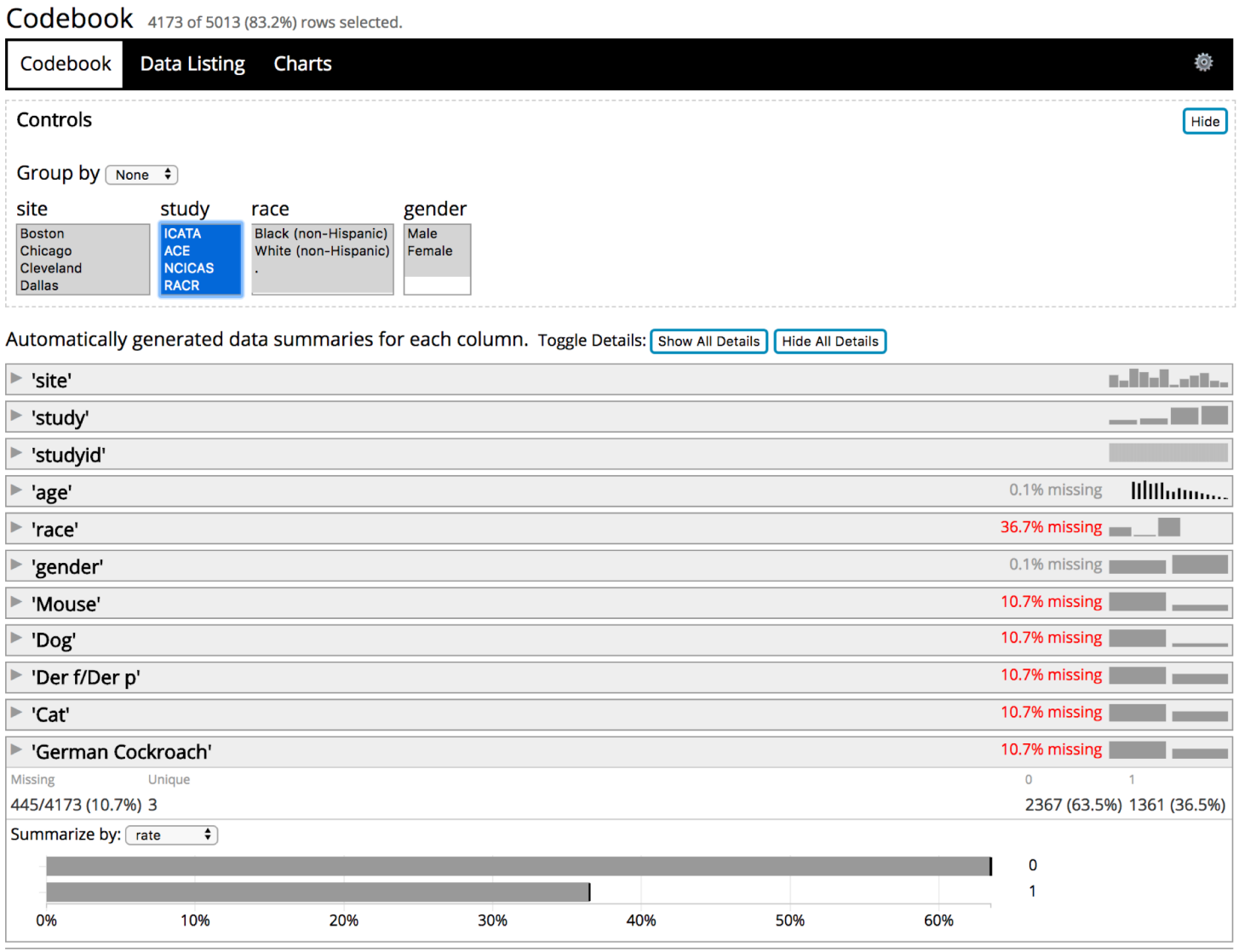
Real-time Interactivity

Filtering, Grouping and Highlighting



Exploratory Data Analysis: Web Codebook

Links: [Live Demo]



Exploratory Data Analysis: Web Codebook

2 days later ...

Request #2: Great! Can you break this down by Site?

Web-codebook workflow: Group by site (or filter or make a chart!)

A week passes …

Request #3: Are there differences by gender?

Web-codebook workflow: Group by gender (or filter or make a chart!)

4 days after that …

Request #4: How many are both mouse and cockroach allergic?

Web-codebook workflow: Highlight cockroach allergic participants (of filter or group on allergic status!).

2 months later while you're on vacation …

Request #5: That paper is due tomorrow! Can I have p-values for all of that??

Web Codebook - Data Requirements

- Creates data summary for any tabular data set
- No configuration, data specification or data mapping needed
- Support for metadata via 2nd data set
- Tool uses JSON data
  - Many functions available to map from other formats
  - R implementation handles this automatically via the rio package
- Lots of configurable options
  - Tool picks smart defaults for filters and groups
  - User can customize settings in real time

Web Codebook – Technical Details

- Written in Javascript
- Works in any modern web browser
- Easy to implement in any web environment
- Minimal Dependencies
  - D3.js (Bostock, 2011)
  - webcharts.js (Bryant, 2016)
- In most cases, a summary for a single data set can be initialized with a single line of javascript:

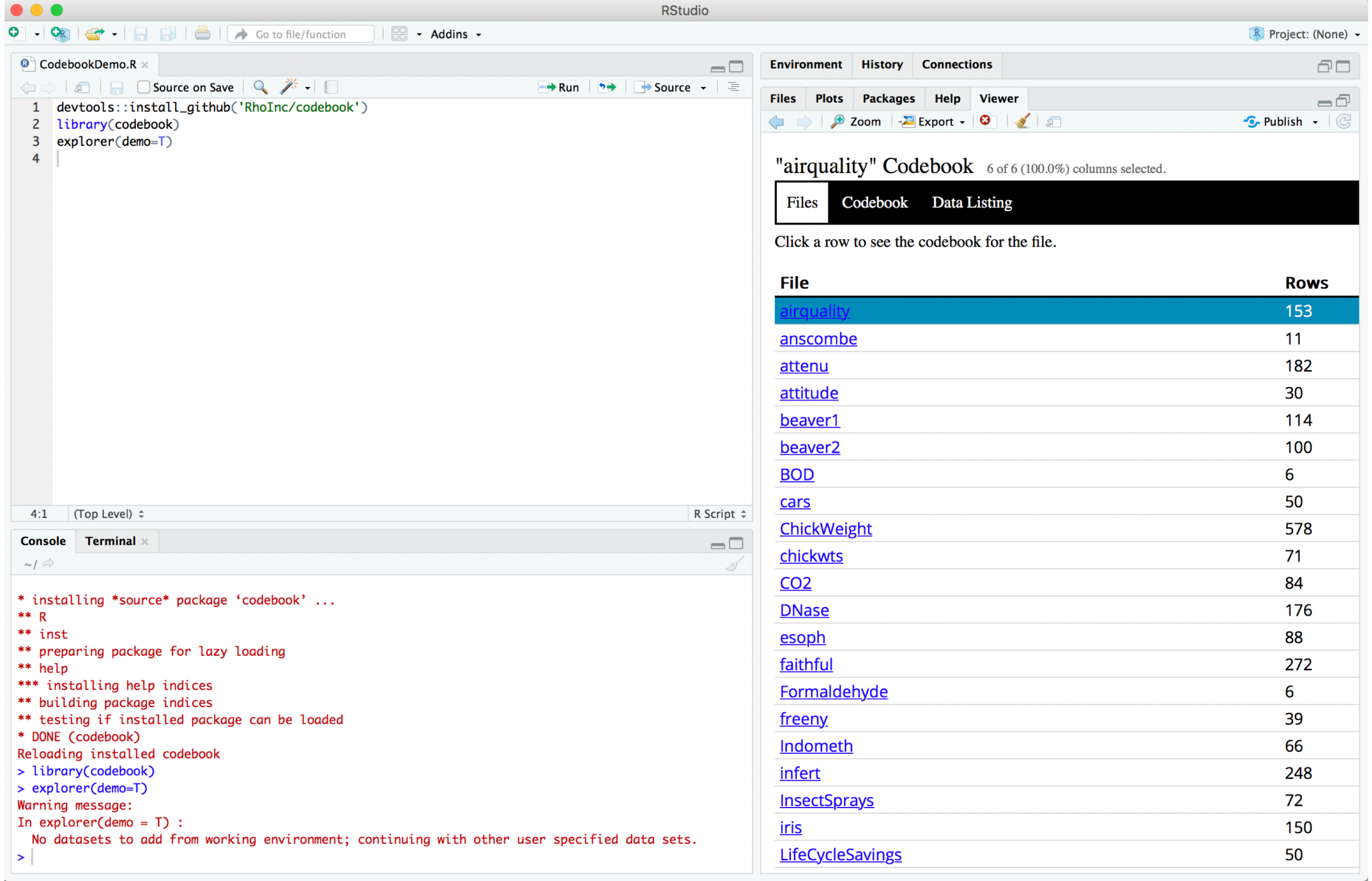
```
webcodebook.createChart('#chartLocation', {}).init(data)
```

- Full technical documentation in project wiki

R Implementation – Technical Details

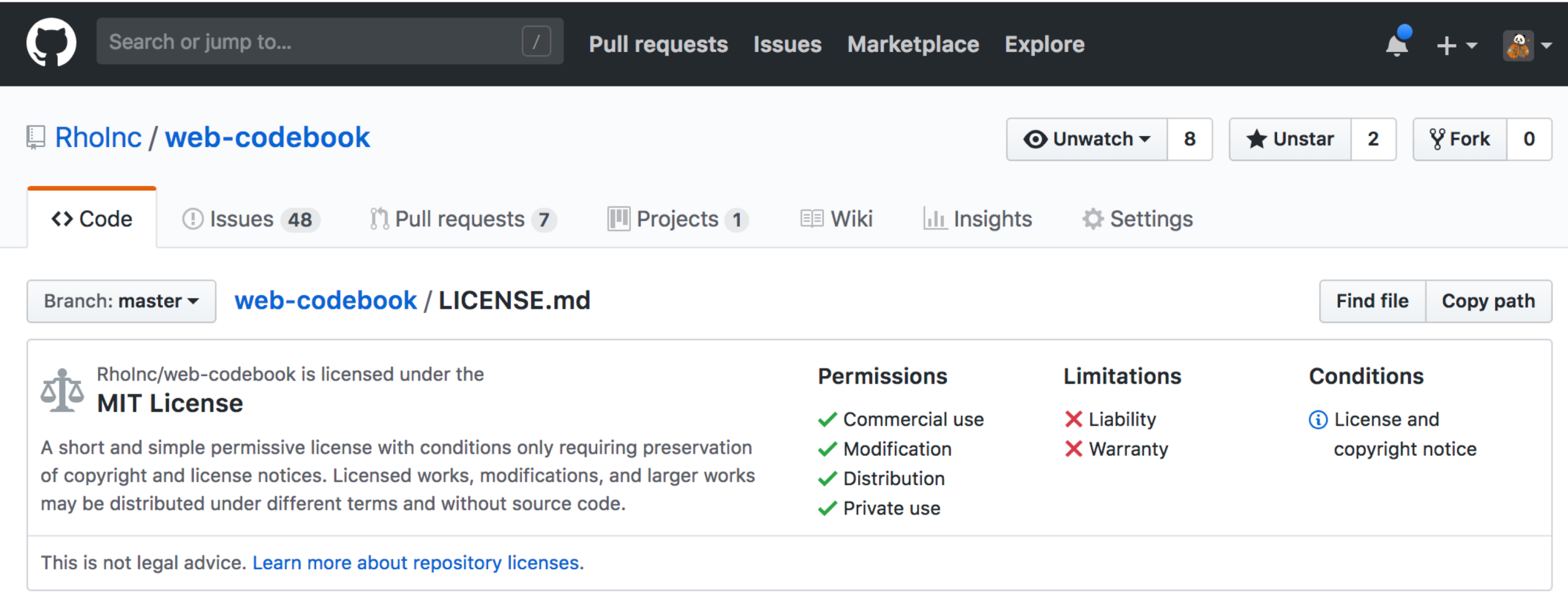
- *codebook()* function provides htmlwidgets wrapper for web codebook
- *explorer()* function loads all data objects in current session by default
  - Set demo=T to load all sample data sets instead
- *codebookApp()* - shiny app lets users load data sets in real time
- All functions create output that is easy to export and share.
- All can be initialized in 1 line of code after loading the library:

```
> devtools::install_github('RhoInc/codebook')
> library(codebook)
> codebook(mtcars) #create webpage with htmlwidgets
> explorer() #create a codebook explorer for current environment
> codebookApp() #run shiny app
```


RStudio
Go to file/function
Addins
Project: (None)
CodebookDemo.R
Source on Save
Run
Source
1 devtools::install_github('RhoInc/codebook')
2 library(codebook)
3 explorer(demo=T)
4
4:1
(Top Level)
R Script
Console
Terminal
~/
* installing *source* package 'codebook' ...
** R
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (codebook)
Reloading installed codebook
> library(codebook)
> explorer(demo=T)
Warning message:
In explorer(demo = T) :
No datasets to add from working environment; continuing with other user specified data sets.
>
Environment
History
Connections
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
"airquality" Codebook
6 of 6 (100.0%) columns selected.
Files
Codebook
Data Listing
Click a row to see the codebook for the file.
File
Rows
airquality 153
anscombe 11
attenu 182
attitude 30
beaver1 114
beaver2 100
BOD 6
cars 50
ChickWeight 578
chickwts 71
CO2 84
DNase 176
esoph 88
faithful 272
Formaldehyde 6
freeny 39
Indometh 66
infert 248
InsectSprays 72
iris 150
LifeCycleSavings 50

Web Codebook – Free and Open Source

Link: [MIT License]



Web Codebook – Free and Open Source

Link: Rho's Open Source Handbook

Rho Open Source Handbook

Overview

This library provides information about open source software development at Rho. We provide a brief overview of Rho's philosophy regarding open source development, and then give guidelines for Rho staff and external contributors working on Rho's open source projects.

Planned updates to this document are tracked as GitHub issues. Questions can be submitted on the issues page or via email.

What is open source?

The Open Source Initiative has a good definition:

> Generally, Open Source software is software that can be freely accessed, used, changed, and shared (in modified or unmodified form) by anyone.

What is Rho's position on open source code sharing?

Rho is in favor of open source code sharing. An open source approach takes the work done at Rho and broadens its reach. That's a great thing since it falls in line with our core purpose:

> To improve health, extend life, and enhance quality of life through corporate and research excellence.

Next Steps

- Continue to improve interactive data summaries (issues tracker)
- Release v1.0 of R package on CRAN
- Add statistical testing via R in v2.0 of R package!