Hilary Mason
@hmason

Following

Data people: What is the very first thing you do when you get your hands on a new data set?

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
43	78

9:56 PM - 11 Jun 2014

Reply to @hmason

Iris Classon @IrisClasson · Jun 11
@hmason look for people I know? Just kidding :)
Details Reply Retweet Favorite ••• More

Javier Moreno @infrahumano · Jun 11
@hmason: clean it up.
Details Reply Retweet Favorite ••• More

Cam Davidson-Pilon @Cmrn_DP · Jun 11
@hmason Look for None/NaNs
Details Reply Retweet Favorite ••• More

Dr. Jennie Chen-ergy @MisoHungry · Jun 11
@hmason Depends, do I already know the structure of the data?
Details Reply Retweet Favorite More

Raymond Lilly @37point2 · Jun 11
@hmason Get overly excited and think about things I can do with it.
Details Reply Retweet Favorite More

Andre Bach @Nonnormalizable · Jun 11
@hmason Check if the "primary key" is in fact the primary key.
Details Reply Retweet Favorite More

Francisco Alvarez @FrankAlvarez · Jun 11
@hmason Depends on the format, but usually check for missing values/NaNs.
Details Reply Retweet Favorite More

chuck b. @back40feet · Jun 11
@hmason Count rows and columns. Determine if it wide or long. Assess presence of missing values, continuous variables, discrete variables.
Details Reply Retweet Favorite More

Chris Rorie @chrisrorie · Jun 11
@hmason visualize it, charts are critical (for me at least)
Details Reply Retweet Favorite More

Dan Kozikowski @dfkoz · Jun 11
@hmason a thorough cleaning
Details Reply Retweet Favorite More

Evan Hensleigh @futuraprime · Jun 11
@hmason Read the dataset's documentation.
Details Reply Retweet Favorite More

Aaron MacNeil @ma_macneil · Jun 11
@hmason check spellings are consistent; then Cleveland plots
Details Reply Retweet Favorite More

sgoggins @sgoggins · Jun 11
@hmason descriptive statistics, ID missing data & preliminary notes on what the data *is* .. Accept nothing at face value. :)
Details Reply Retweet Favorite More

Billy Sharp @dqchronicle · Jun 11
@hmason distribution graphs and/or counts
Details Reply Retweet Favorite More

Tommy Levi @tslevi · Jun 11
@hmason str(data) and summary(data). Tells me what I'm dealing with and basic summary stats. Next is check for NA and nulls.
Details Reply Retweet Favorite More

brian abelson @brianabelson · Jun 11
@hmason try to get my hands on the docs!
Details Reply Retweet Favorite More

Herbie Lewis @HerbieLewis · Jun 11
@hmason Figure out what I want to do with it, what I want to use it for, and how the insights from it will fit into my greater task or goals
Details Reply Retweet Favorite More

Jonathan Robinson @jon_m_rob · Jun 11
.@hmason Obviously the first thing I do when I get a new dataset is fit some backpropogated convolutional neural nets :p
Details Reply Retweet Favorite More

Justin Lintz @lintzston · Jun 11
@hmason delete it, oh you said data people
Details Reply Retweet Favorite More

Adam Laiacano @adamlaiacano · Jun 11
@hmason summary()
Details Reply Retweet Favorite More

Arek Stopczynski @h0pbeat · Jun 11
@hmason histogram everything.
Details Reply Retweet Favorite More

Hilary Mason @hmason · Jun 11
@lintzston ha!
Details Reply Retweet Favorite More

Pete Warden @petewarden · Jun 11
@hmason yep, been looking through the other answers, there's lots of interesting stuff!
Details Reply Retweet Favorite More

Josh Montague @jrmontag · Jun 11
@hmason Would love for you to tally/share the responses!
Details Reply Retweet Favorite More

Carol Davidsen @cld276 · Jun 11
@hmason run a bunch of count/groupby statements to gauge if I think it's corrupt.
Details Reply Retweet Favorite More

Jeroen Janssens @jeroenhjanssens · Jun 11
@hmason $ ls -1sh; wc -l; head; csvlook || jq; ...
Details Reply Retweet Favorite More

Hilary Mason @hmason · Jun 11
@jrmontag I'll try, but there are so many coming in I can barely keep up!
Details Reply Retweet Favorite More

Mark Huckabee @mhuckabee1 · Jun 11
@hmason Load it in R and run descriptive stats.
Details Reply Retweet Favorite More

Alexander Furnas @zfurnas · Jun 11
@hmason read data dictionary/documentation. If it is already tabular, I look at distributions, then scatterplot matrix of interesting vars.
Details Reply Retweet Favorite More

Stuart Robinson @stuartrobinson · Jun 11
@hmason Salivate
Details Reply Retweet Favorite More

Josh Montague @jrmontag · Jun 11
@hmason If you can wait until tomorrow morning, I can probably help :)
Details Reply Retweet Favorite More

Derek Willis @derekwillis · Jun 11
@cld276 @hmason sorting to find out where the blanks and nulls are, too.
Details Reply Retweet Favorite More

Alan Williams @alanwilliams1 · Jun 11
@petewarden @hmason 'ls -lah', 'head ./*', 'wc -l ./*', back up a raw copy locally and to S3, 'ipython notebook --ip="*"'
Details Reply Retweet Favorite More

Tariq Khokhar @tkb · Jun 11
.@hmason usually less & csvkit to get a feel for it, sed / Open Refine for tidy up. And naturally, 3D-print / laser cut a plot of it.
Details Reply Retweet Favorite More

Hilary Mason @hmason · Jun 11
@jrmontag That'd be awesome! I'm also on airplane wifi right now which is now helping.
Details Reply Retweet Favorite More

nick trendov @ManyCUES · Jun 11
easy least pumpkin pie @hmason Strip away as much as possible
Details Reply Retweet Favorite More

Fred Benenson @fredbenenson · Jun 11
@hmason wc -l
Details Reply Retweet Favorite More

Hilary Mason @hmason · Jun 11

Jacob @japerk · Jun 11
@hmason figure out the format & how to read it. Then ask myself, what can be learned from this data?
Details Reply Retweet Favorite More

Data Pointed @DataPointed · Jun 11
@hmason assess its integrity
Details Reply Retweet Favorite More

Erin Jonaitis @emjonaitis · Jun 11
@hmason Scan it for weirdness -- ragged edges, errant data types, implausible values, nonstandard format.
Details Reply Retweet Favorite More

Dan Couture @MathYourLife · Jun 11
@hmason check the size. It'll influence analyses used, resources required, visualizations, if random subsets are required for spelunking...
Details Reply Retweet Favorite More

Michael Griffiths @msjgriffiths · Jun 11
@hmason Check for missing data. Check max/min values. Check units are consistent. Check data types (float, int, string, time stamp, etc).
Details Reply Retweet Favorite More

.~* mArC *~. @slpsys · Jun 11
@hmason I put on my regex and wizard hat.
Details Reply Retweet Favorite More

Susan Dynarski @dynarski · Jun 11
@hmason @erikbryn Figure out how they code missing & non-response lest I make inferences about a variable missing for 70% of the sample.
Details Reply Retweet Favorite More

Tariq Khokhar @tkb · Jun 11
@hmason 3D/Laser plots not a regular occurrence sadly (get them mail ordered) - but some nice histograms etc. in acrylic and MDF!
Details Reply Retweet Favorite More