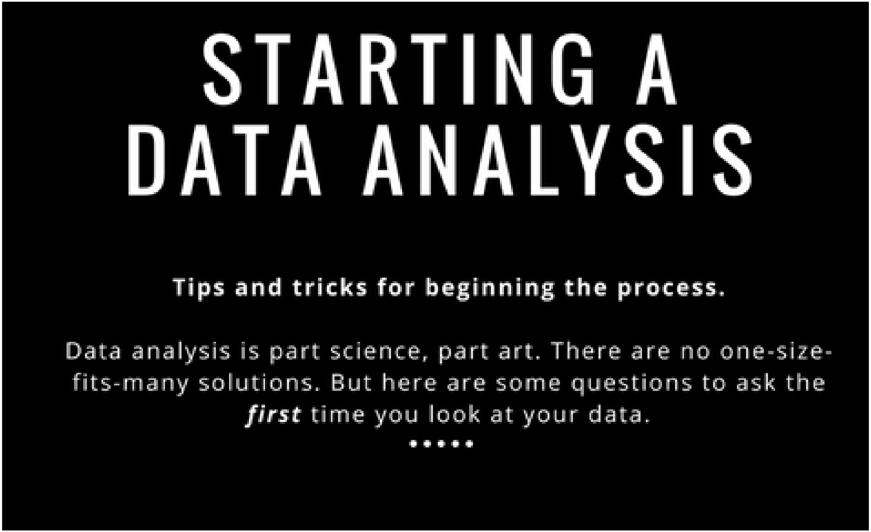
Data, Data, Data - Pt 2.

Dr. Matthew Sigal[1]

Feb. 26th, 2021

[1] . . . with many wonderful illustrations from Allison Horst!

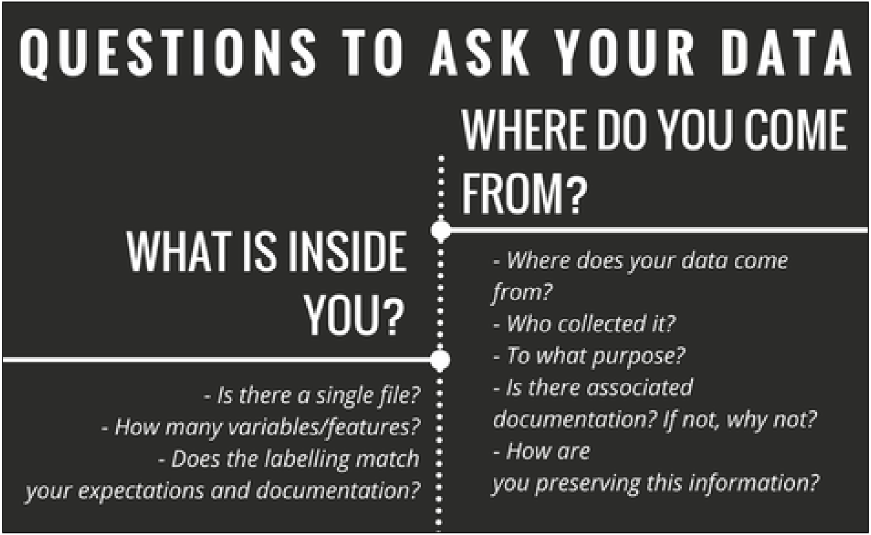


STARTING A DATA ANALYSIS

Tips and tricks for beginning the process.

Data analysis is part science, part art. There are no one-size-fits-many solutions. But here are some questions to ask the ***first*** time you look at your data.

•••••



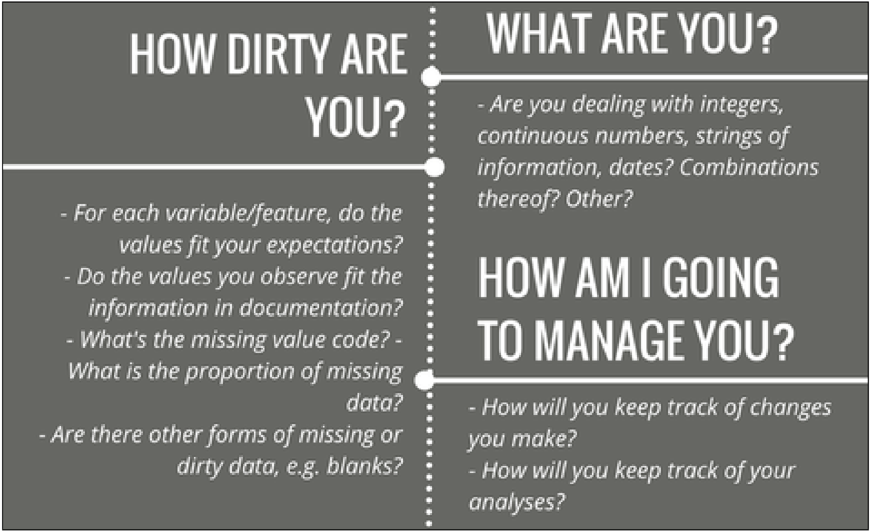
QUESTIONS TO ASK YOUR DATA

WHAT IS INSIDE YOU?

- Is there a single file?
- How many variables/features?
- Does the labelling match your expectations and documentation?

WHERE DO YOU COME FROM?

- Where does your data come from?
- Who collected it?
- To what purpose?
- Is there associated documentation? If not, why not?
- How are you preserving this information?



HOW DIRTY ARE YOU?

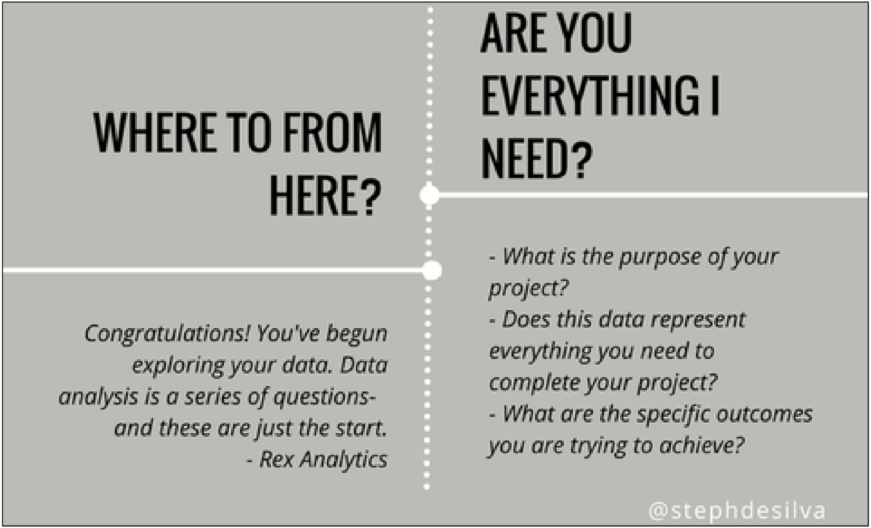
- For each variable/feature, do the values fit your expectations?
- Do the values you observe fit the information in documentation?
- What's the missing value code? - What is the proportion of missing data?
- Are there other forms of missing or dirty data, e.g. blanks?

WHAT ARE YOU?

- Are you dealing with integers, continuous numbers, strings of information, dates? Combinations thereof? Other?

HOW AM I GOING TO MANAGE YOU?

- How will you keep track of changes you make?
- How will you keep track of your analyses?



WHERE TO FROM HERE?

Congratulations! You've begun exploring your data. Data analysis is a series of questions- and these are just the start.
- Rex Analytics

ARE YOU EVERYTHING I NEED?

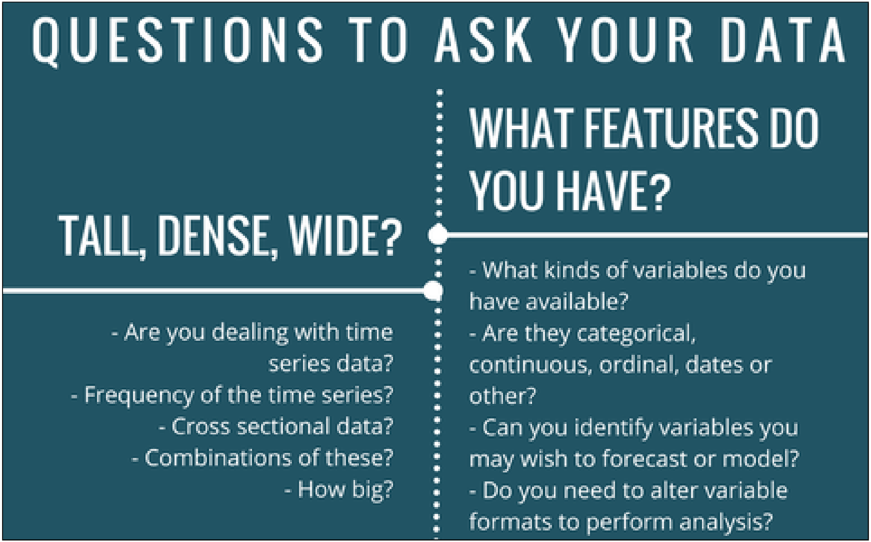
- What is the purpose of your project?
- Does this data represent everything you need to complete your project?
- What are the specific outcomes you are trying to achieve?

@stephdesilva

DATA ANALYSIS: A NEXT STEP

Tips and tricks for developing your analysis process.

Data analysis is part science, part art. There is no simple recipe for nuanced understanding of your data. Here are some questions to ask yourself as you continue to look at your data.



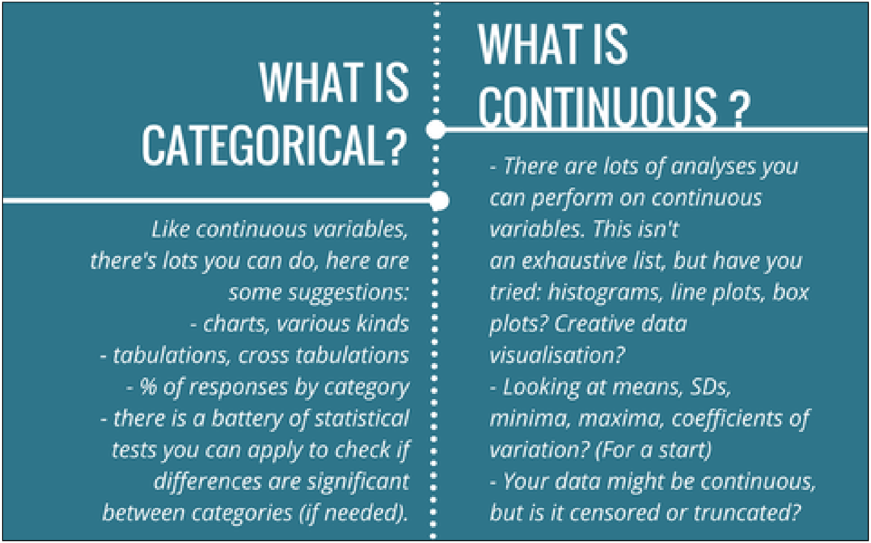
QUESTIONS TO ASK YOUR DATA

TALL, DENSE, WIDE?

- Are you dealing with time series data?
- Frequency of the time series?
- Cross sectional data?
- Combinations of these?
- How big?

WHAT FEATURES DO YOU HAVE?

- What kinds of variables do you have available?
- Are they categorical, continuous, ordinal, dates or other?
- Can you identify variables you may wish to forecast or model?
- Do you need to alter variable formats to perform analysis?



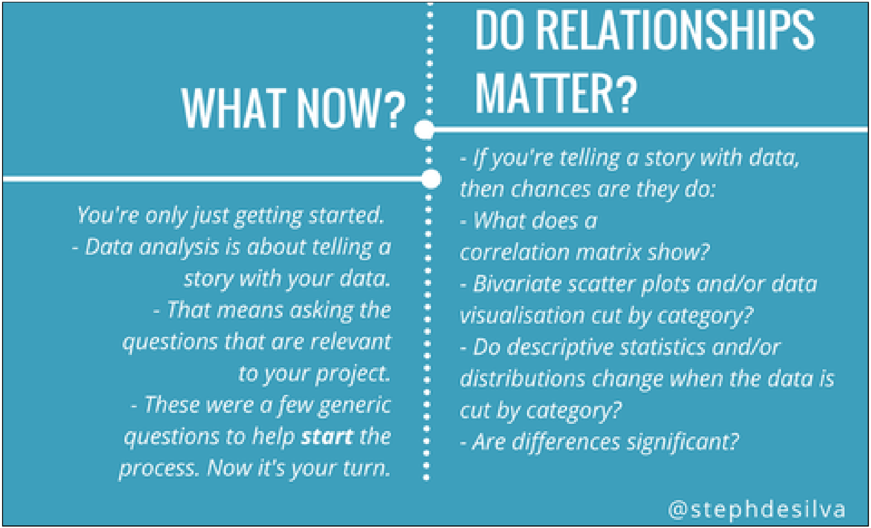
WHAT IS CATEGORICAL?

Like continuous variables, there's lots you can do, here are some suggestions:

- charts, various kinds
- tabulations, cross tabulations
- % of responses by category
- there is a battery of statistical tests you can apply to check if differences are significant between categories (if needed).

WHAT IS CONTINUOUS ?

- There are lots of analyses you can perform on continuous variables. This isn't an exhaustive list, but have you tried: histograms, line plots, box plots? Creative data visualisation?
- Looking at means, SDs, minima, maxima, coefficients of variation? (For a start)
- Your data might be continuous, but is it censored or truncated?



WHAT NOW?

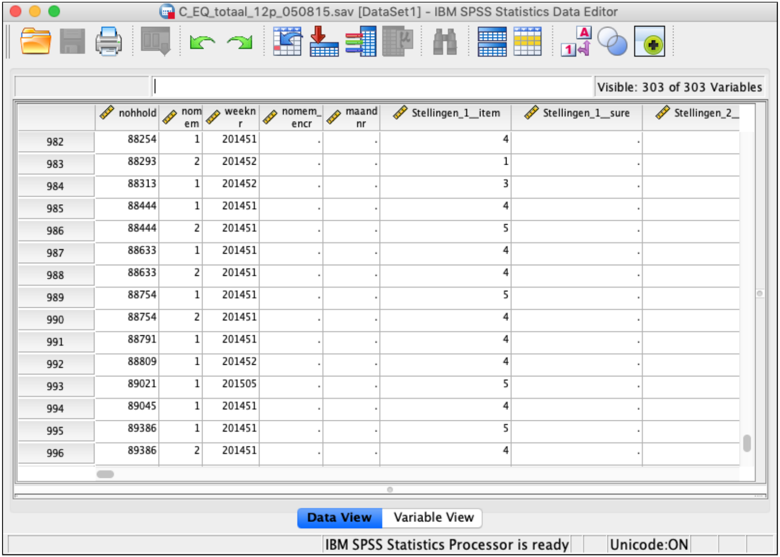
You're only just getting started.
- Data analysis is about telling a story with your data.
- That means asking the questions that are relevant to your project.
- These were a few generic questions to help **start** the process. Now it's your turn.

DO RELATIONSHIPS MATTER?

- If you're telling a story with data, then chances are they do:
- What does a correlation matrix show?
- Bivariate scatter plots and/or data visualisation cut by category?
- Do descriptive statistics and/or distributions change when the data is cut by category?
- Are differences significant?

@stephdesilva

Real Life Example – The EQ-i:



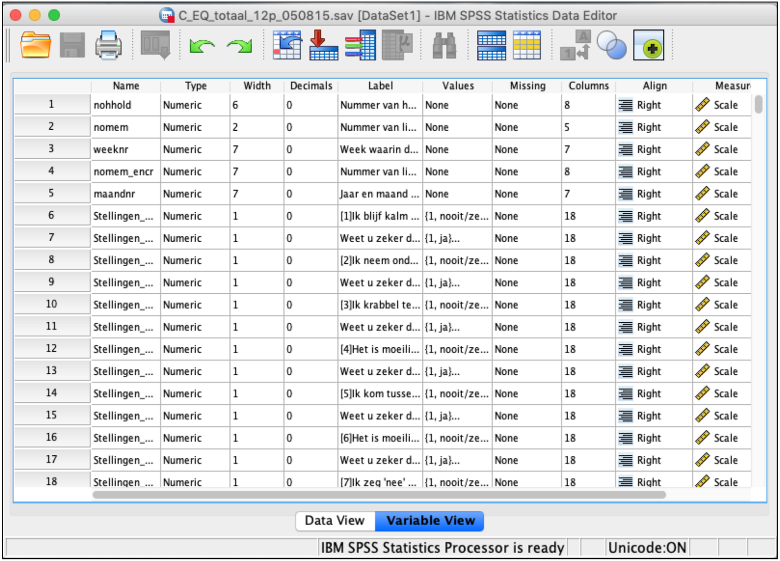
C_EQ_totaal_12p_050815.sav [DataSet1] - IBM SPSS Statistics Data Editor

Visible: 303 of 303 Variables

	nohhold	nomem	weeknr	nomem_encr	maandnr	Stellingen_1__item	Stellingen_1__sure	Stellingen_2_
982	88254	1	201451	.	.	4	.	
983	88293	2	201452	.	.	1	.	
984	88313	1	201452	.	.	3	.	
985	88444	1	201451	.	.	4	.	
986	88444	2	201451	.	.	5	.	
987	88633	1	201451	.	.	4	.	
988	88633	2	201451	.	.	4	.	
989	88754	1	201451	.	.	5	.	
990	88754	2	201451	.	.	4	.	
991	88791	1	201451	.	.	4	.	
992	88809	1	201452	.	.	4	.	
993	89021	1	201505	.	.	5	.	
994	89045	1	201451	.	.	4	.	
995	89386	1	201451	.	.	5	.	
996	89386	2	201451	.	.	4	.	

IBM SPSS Statistics Processor is ready

Unicode:ON

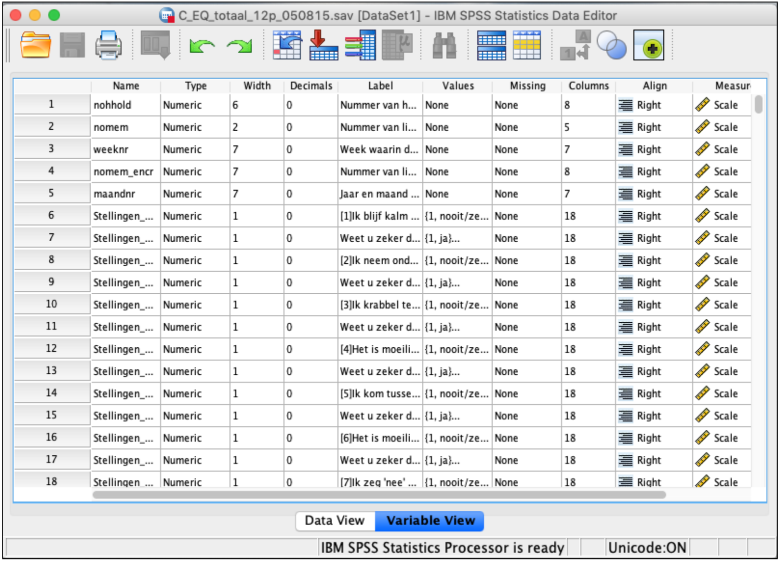


C_EQ_totaal_12p_050815.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measur
1	nohhold	Numeric	6	0	Nummer van h...	None	None	8	Right	Scale
2	nomem	Numeric	2	0	Nummer van li...	None	None	5	Right	Scale
3	weeknr	Numeric	7	0	Week waarin d...	None	None	7	Right	Scale
4	nomem_encr	Numeric	7	0	Nummer van li...	None	None	8	Right	Scale
5	maandnr	Numeric	7	0	Jaar en maand ...	None	None	7	Right	Scale
6	Stellingen_...	Numeric	1	0	[1]Ik blijf kalm ...	{1, nooit/ze...	None	18	Right	Scale
7	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
8	Stellingen_...	Numeric	1	0	[2]Ik neem ond...	{1, nooit/ze...	None	18	Right	Scale
9	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
10	Stellingen_...	Numeric	1	0	[3]Ik krabbel te...	{1, nooit/ze...	None	18	Right	Scale
11	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
12	Stellingen_...	Numeric	1	0	[4]Het is moeili...	{1, nooit/ze...	None	18	Right	Scale
13	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
14	Stellingen_...	Numeric	1	0	[5]Ik kom tusse...	{1, nooit/ze...	None	18	Right	Scale
15	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
16	Stellingen_...	Numeric	1	0	[6]Het is moeili...	{1, nooit/ze...	None	18	Right	Scale
17	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
18	Stellingen_...	Numeric	1	0	[7]Ik zeg 'nee' ...	{1, nooit/ze...	None	18	Right	Scale

IBM SPSS Statistics Processor is ready

Unicode:ON



C_EQ_totaal_12p_050815.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measur
1	nohhold	Numeric	6	0	Nummer van h...	None	None	8	Right	Scale
2	nomem	Numeric	2	0	Nummer van li...	None	None	5	Right	Scale
3	weeknr	Numeric	7	0	Week waarin d...	None	None	7	Right	Scale
4	nomem_encr	Numeric	7	0	Nummer van li...	None	None	8	Right	Scale
5	maandnr	Numeric	7	0	Jaar en maand ...	None	None	7	Right	Scale
6	Stellingen_...	Numeric	1	0	[1]Ik blijf kalm ...	{1, nooit/ze...	None	18	Right	Scale
7	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
8	Stellingen_...	Numeric	1	0	[2]Ik neem ond...	{1, nooit/ze...	None	18	Right	Scale
9	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
10	Stellingen_...	Numeric	1	0	[3]Ik krabbel te...	{1, nooit/ze...	None	18	Right	Scale
11	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
12	Stellingen_...	Numeric	1	0	[4]Het is moeili...	{1, nooit/ze...	None	18	Right	Scale
13	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
14	Stellingen_...	Numeric	1	0	[5]Ik kom tusse...	{1, nooit/ze...	None	18	Right	Scale
15	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
16	Stellingen_...	Numeric	1	0	[6]Het is moeili...	{1, nooit/ze...	None	18	Right	Scale
17	Stellingen_...	Numeric	1	0	Weet u zeker d...	{1, ja}...	None	18	Right	Scale
18	Stellingen_...	Numeric	1	0	[7]Ik zeg 'nee' ...	{1, nooit/ze...	None	18	Right	Scale

Instead, let's read it in via R. . . [see `EQi-DataImport.pdf`]

Your Data Contract

"Data Organization in Spreadsheets"

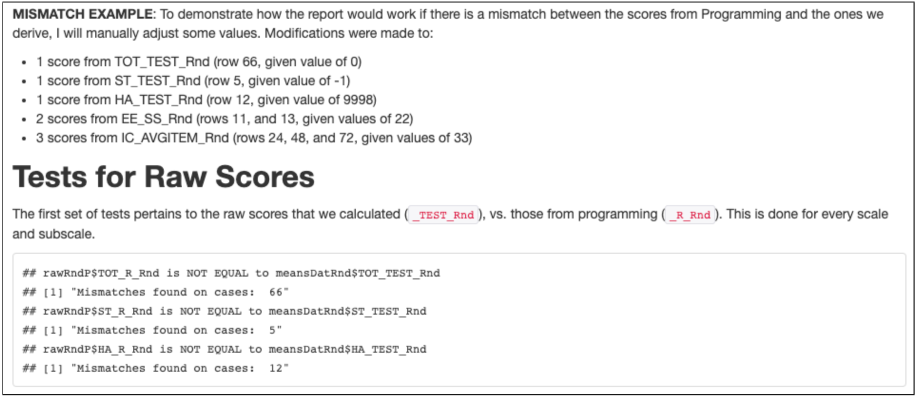
Broman and Woo (2018) in *American Statistician*:
https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

> *Spreadsheets, for all of their mundane rectangularness, have been the subject of angst and controversy for decades. Some writers have admonished that "real programmers don't use spreadsheets" and that we must "stop that subversive spreadsheet."*

Broman & Woo's Rules to Live By

1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Make it a rectangle
7. Create a data dictionary
8. No calculations
9. Do not use font color or highlighting as data
10. Make backups
11. Use data validation [see the `EQ360-Swedish-Testing` pdfs]
12. Save the data in plain text files

Data Validation



MISMATCH EXAMPLE: To demonstrate how the report would work if there is a mismatch between the scores from Programming and the ones we derive, I will manually adjust some values. Modifications were made to:

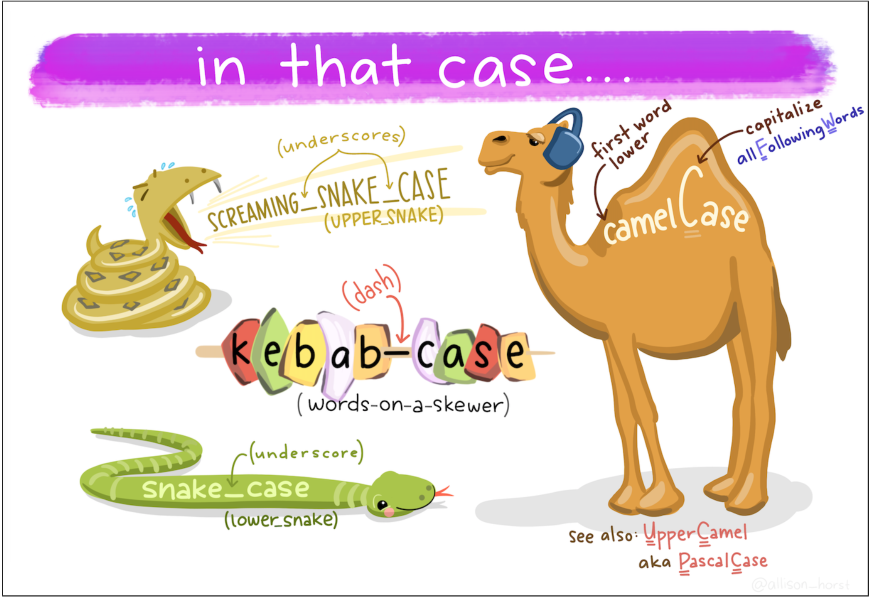
- 1 score from TOT_TEST_Rnd (row 66, given value of 0)
- 1 score from ST_TEST_Rnd (row 5, given value of -1)
- 1 score from HA_TEST_Rnd (row 12, given value of 9998)
- 2 scores from EE_SS_Rnd (rows 11, and 13, given values of 22)
- 3 scores from IC_AVGITEM_Rnd (rows 24, 48, and 72, given values of 33)

Tests for Raw Scores

The first set of tests pertains to the raw scores that we calculated (`_TEST_Rnd`), vs. those from programming (`_R_Rnd`). This is done for every scale and subscale.

```
## rawRndP$TOT_R_Rnd is NOT EQUAL to meansDatRnd$TOT_TEST_Rnd
## [1] "Mismatches found on cases:  66"
## rawRndP$ST_R_Rnd is NOT EQUAL to meansDatRnd$ST_TEST_Rnd
## [1] "Mismatches found on cases:  5"
## rawRndP$HA_R_Rnd is NOT EQUAL to meansDatRnd$HA_TEST_Rnd
## [1] "Mismatches found on cases:  12"
```

Variable Names



Variable Names

janitor::clean_names
convert all column names to _*_case!
not so awesome column names
FUN(units)!
Sample2.2!
Such.?Mess!
ooo.OOF-.
MESS
Mm.,OOP
clean_names(your data frame + case choice here)
CHOOSE:
✓ snake
lower_camel
big_camel
screaming_snake
+ more!
WAY MORE DEALWITHABLE COLUMN NAMES!
so_much_nicer
_yay_ya_
hope_rest
_wo
ooo_
_cool
teddy
_than
good_feel_
_a_
snake
HORST 2019

For a discussion on the choice of variable names and using `pointblank` for data checks, see Emily Riederer's post “Column Names as Contracts” (https://emilyriederer.netlify.app/post/column-name-contracts/).