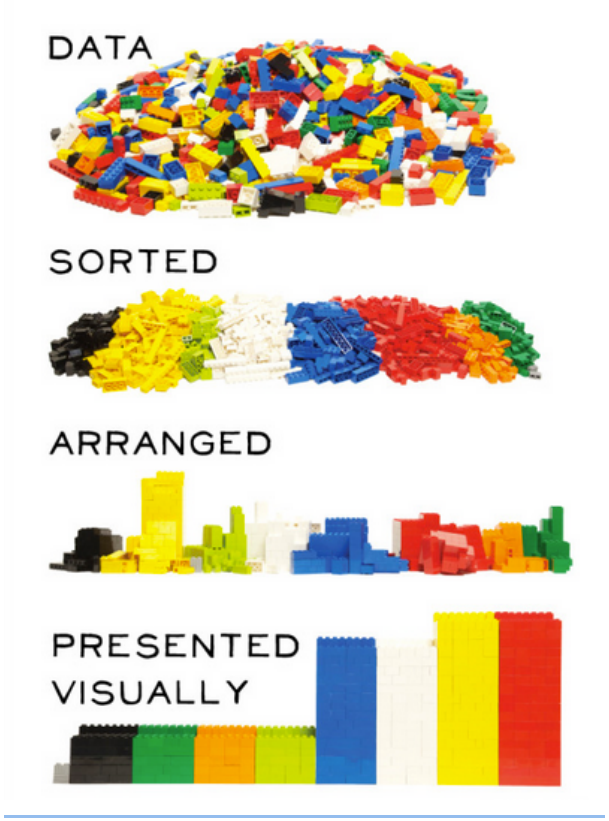
Basements of Big Data analysis

Simple but worthy statements beneath the data cleaning

Hanna Rudakouskaya, Data Analyst, Teqniksoft

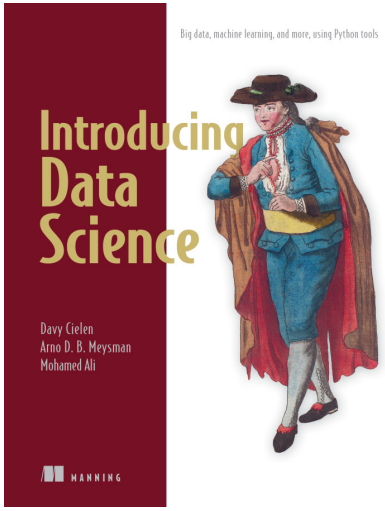
June 27, 2017

Inspiration: garbage in - garbage out: nothing is more true!



"Sponsors" of today's theme

- Davy Cielen, Arno D.B. Meysman, Mohamed Ali "Introducing Data Science: Big Data, Machine Learning, and more using Python tools"
- my colleagues Anton and two Kates and the tasks we solve daily
- Kaggle as antipode



Tegniksoft

kaggle

Practice follows: example datasets

1. Dataset with "some" test results: contains 13564 rows and 1854 columns.

2. Datasets with test times for these test results:

- first process: contains 1381088 rows and 43 columns;
- second process: contains 869344 rows and 43 columns.

Quick summary on the data:

```
> head(summary(testtime_64_process))
 header2_col_1      header2_col_2      header2_col_3
 Length:1381088     Length:1381088     Min.   :2017-02-01 20:09:12
 Class :character   Class :character   1st Qu.:2017-02-07 01:34:55
 Mode  :character   Mode  :character   Median :2017-02-09 14:37:07
                                       Mean   :2017-02-14 02:47:00
                                       3rd Qu.:2017-02-15 10:56:25
                                       Max.   :2017-04-20 18:20:39
 header2_col_4 header2_col_5 header2_col_9   test_result2_col_19
 Min.   :64    Min.   :0     Min.   :6.000   Min.   :      0
 1st Qu.:64    1st Qu.:0     1st Qu.:6.000   1st Qu.:      0
 Median :64    Median :0     Median :6.000   Median :   2468
 Mean   :64    Mean   :0     Mean   :6.869   Mean   : 274937
 3rd Qu.:64    3rd Qu.:0     3rd Qu.:8.000   3rd Qu.:  38804
 Max.   :64    Max.   :0     Max.   :8.000   Max.   :4358016
```

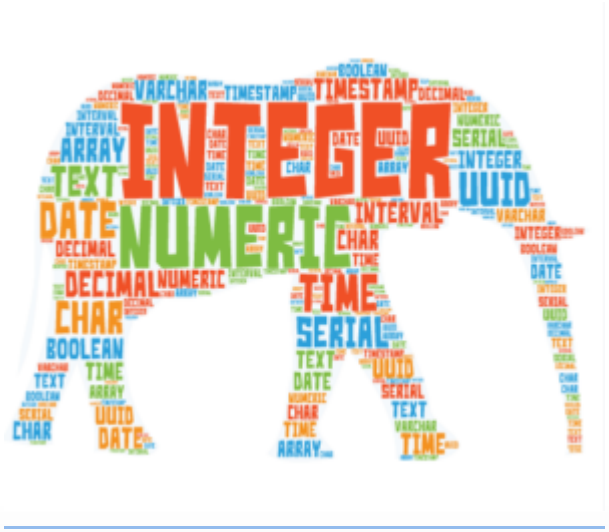
Must-do

obligatory "sanity checks" when starting DA process

Data Types

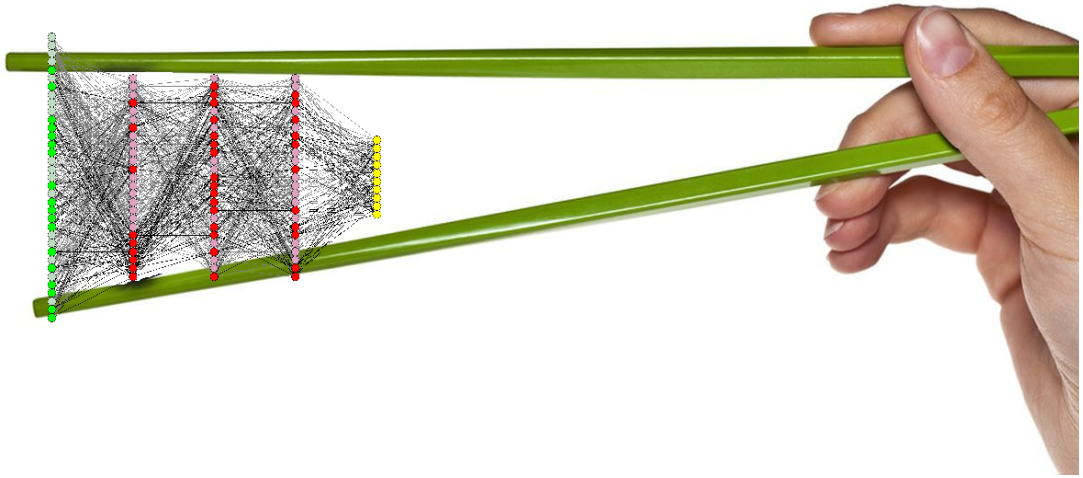
Always check data types before starting anything else. Most common points:

1. does string variable really mean string? Or factor?
2. are all the numerical data columns factors?
3. are all factors just factors or some of them are ranked?



Unofficial rule: if unique values are 3% or less of all the data length - that might be a factor

Both Python, R (and Excel :0) are trying to guess the column types - don't forget to cross-check it! Sometimes reading-all-in as characters and transforming onto numeric is better.



Certain methods "eat" certain types and you'd better not to cheat on it!

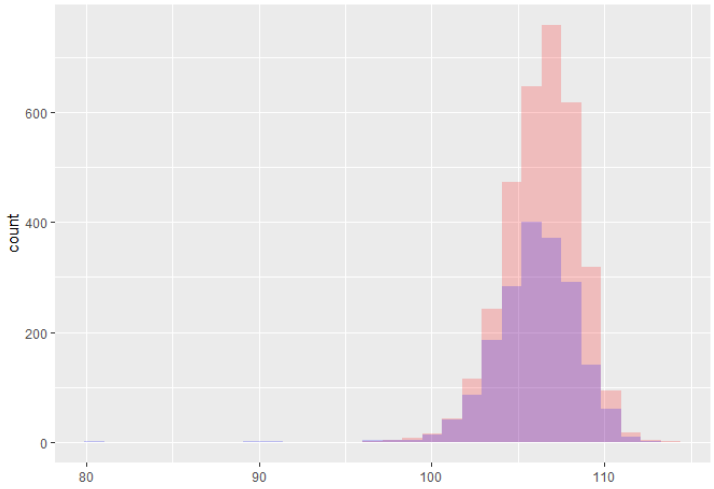
Data types example:

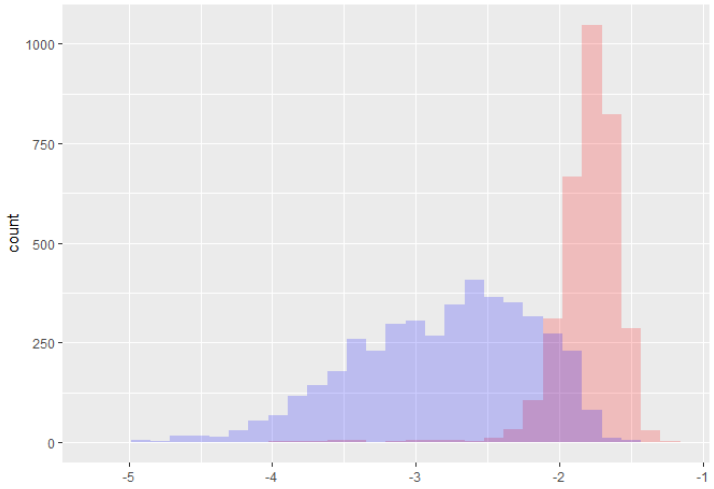
DEMO:

`header_col_3` (POSIX to integer)	`test_result_col_1700` (factor to integer)	`test_result_col_1282` (double as-is)
Min.: "2017-02-01 20:09:12"	-35: 3400	Min.: 0.0000
1st Qu.: "2017-02-07 01:35:31"	-17: 10156	1st Qu.: 0.7165
Median: "2017-02-09 14:31:11"		Median: 1.1360
Mean: "2017-02-13 10:07:59"		Mean: 1.2220
3rd Qu.: "2017-02-13 03:57:39"		3rd Qu.: 1.4560
Max.: "2017-04-13 09:29:43"		Max.: 57.2200
NA's: 0		NA's: 530

Data distributions

- features distributions expected to be fixed;
- check it when you receive new data to apply your pre-built model:





NA values

- Is it a good idea to omit NAs?

see what will happen if we'll omit NAs from our data sets directly: rows in data with NA omit directly is **0**!

So we'd better deal with NA more intelligently ;)

Hint: if there are lots of NAs in dataset, make a separate variable as NA percentage counter for every row - this may help later.

Hint 2: note that header rows rarely have NAs inside them.

Step-by-step NAs omitting:

- remove columns with high percentage of NAs (say, more than 95% of column are NAs);
- remove **rows** with high percentage of NAs (say, more than 90% of row are NAs);
- finally (though you can continue as long as needed!) filter out more columns (say, which have more than 75% of NAs).

What we have how? **12639** rows that are complete! We can omit non-complete rows now.

Hint: block NA test.

Multicollinearity

Inspiration: Mercedes EDA & XGBoost Starter (~0.55) by anokas

- **Constant features**: "Interestingly, we have … features which only have a single value in them - these are pretty useless for supervised algorithms, and should probably be dropped (unless you want to use them for anomaly detection in case a different value appears in the test set)" => filtering out factors with 1 level only

NOTE: some of the header rows have been filtered out as well.

- "Near zero variance": a feature has near zero variance ifit has very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large.

- **Multicollinearity** by itself

NOTE: when removing one of the collinear variables one should always remember that 99% becomes 96% when going through the columns etc.

- **Non-multicollinearity** at first sight: correlation is 0.6

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	-2	-2	-1,996781349	-1,98831749	-1,988079071	-1,987363815	-1,985337257	-1,984622002	-1,984502792	-1,984383583	85,29044342	198,9406433	206,8881226
2	-1,01263618	0	0	0	0	0	0	0	8	45	0	0	0
3	-1,01251698	0	0	0	0	0	0	137	0	0	0	0	0
4	-1,01239777	0	0	0	0	0	0	61	0	0	0	0	0
5	-1,01192093	0	0	0	0	0	116	0	0	0	0	0	0
6	-1,01180172	0	0	0	0	286	0	0	0	0	0	0	0
7	-1,00369549	0	0	124	41	0	0	0	0	0	0	0	0
8	-1,00298023	0	1	0	0	0	0	0	0	0	0	0	0
9	-1	1718	0	0	0	0	0	0	0	0	0	0	0
10	1	1748	0	0	0	0	0	0	0	0	0	0	0
11	1,003695488	0	0	7	0	0	0	0	0	0	0	0	0
12	1,003933907	0	0	0	17	0	0	0	0	0	0	0	0
13	1,01180172	0	0	0	0	0	269	0	0	0	0	0	0
14	1,012397766	0	0	0	0	0	0	38	0	0	0	0	0
15	1,012636185	0	0	0	0	0	0	0	0	211	0	0	0
16	9,24529171	0	0	0	0	0	0	0	0	0	1	0	0
17	14,14021111	0	0	0	0	0	0	0	0	0	0	1	0
18	14,3718195	0	0	0	0	0	0	0	0	0	0	0	1

Outliers

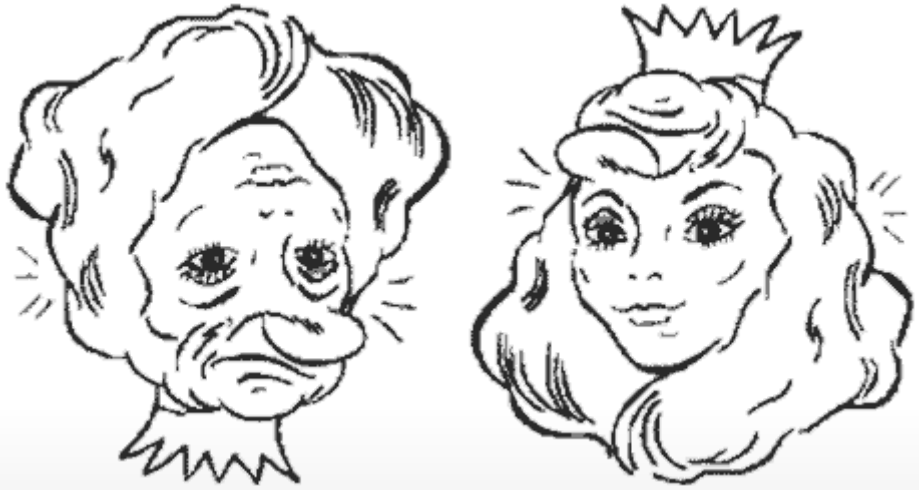
Example rules:

- **mean $\pm$ 3 sigma** — 99.7% of the population is contained within 3 standard deviations from the mean (in case of normal distribution);
- **1Q - 1.5 IQR to 3Q + 1.5 IQR** — Motivation to use this criterion by Rob J. Hyndman;
- **0.01 quantile - 1.5 IQR to 0.99 quantile + 1.5 IQR** — In case we want to remove completely outstanding observations.

Other ideas for preliminary check

- misspellings;
- extra whitespaces;
- whitespace and tab problem;
- lowercase / uppercase.

RegExps are the Dirty Data's worst enemies and Tidy Data's best friends!



Tidying data

Tidy data definition for today

As you probably've heard, according to Hadley Wickham's article "Tidy Data",

"In tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table."

In practice, I'd recommend you to use header analysis beforehand.

Headers and their analysis

First of all, get "table" of header rows meanings:

Some of the header rows are likely to be factors, look at them carefully (`header_col_7` vs. `header_col_10`):

Evident grouping!

Data groups

As we've seen, data header contains several groups. We have 2 options:

- add these groups as factors, or
- separate the whole data set into groups and perform separate analysis for every group.

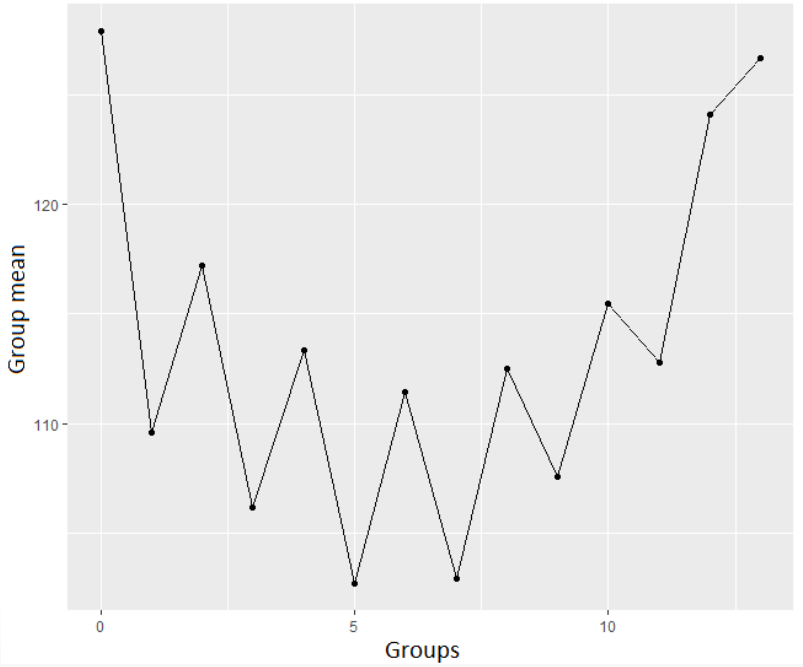
Which option to choose?

1. If you plan to perform classification – option 1;
2. If you plan to perform regression – option 2.

Do I have to have groups at all?

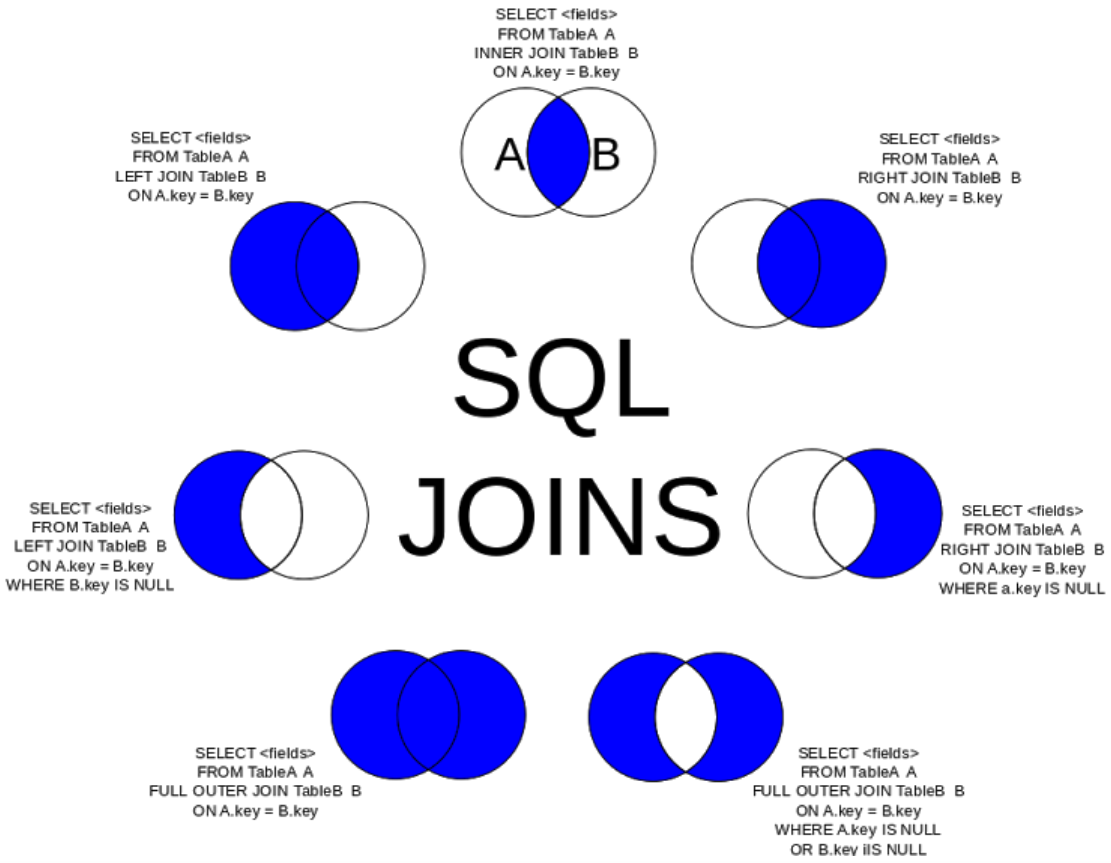
Check:

1. F Test to Compare Variances between groups;
2. Student's t-Test to check significant difference in mean values between the groups.



Data joining

Double-**triple-QUADRUPLE** check what you're doing here!!!



Code book and column specifications

Big Data is usually that big you'll not be ablt to have any code book – just relax and enjoy your non-acquaintance ;)

But! If you're lucky enough to get it - don't waste it!

"Dealing with data" pipeline

Second most important thing in data cleaning.

Don't get lazy - build it!

Solved inside pipeline:

1. Data obtaining (+ updating if necessary);
2. All the data cleaning steps;
3. All the post-cleaning checks (distributions, means, variances etc.).

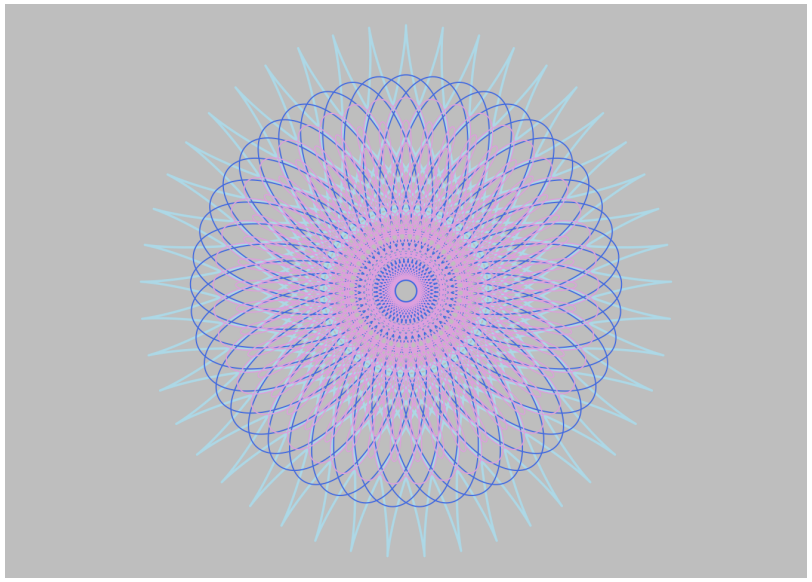
Pipeline must be fixed!

Insight first - analysis later

THE MOST important thing.

- if there's a problem statement - ask for an example;
- if there's no example – feel free to choose any method and play with data in your favourite way! Yummy!
- one dataset may suite one goal and not suite another at all. The goal of analysis must be specific enough so that you can detect this in case.

Enjoy your data!



Hanna Rudakouskaya
Data Analyst, Teqniksoft
http://hannarud.github.io/

Links - FYU

- Davy Cielen, Arno D.B. Meysman, Mohamed Ali "Introducing Data Science: Big Data, Machine Learning, and more using Python tools";
- Teqniksoft Homepage;
- Kaggle: Your Home for Data Science;
- Mercedes EDA & XGBoost Starter (~0.55) by anokas;
- Simple algorithm for online outlier detection of a generic time series;
- "Tidy Data" by Hadley Wickham.